AD_____

AWARD NUMBER: W81XWH-05-1-0204

TITLE: Identification, Characterization and Clinical Development of the New Generation of Breast Cancer Susceptibility Alleles

PRINCIPAL INVESTIGATOR: Nazneen Rahman, M.D., Ph.D.

CONTRACTING ORGANIZATION: The Institute of Cancer Research
London SW7 3RP; United Kingdom

REPORT DATE: March 2011

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1 March 2011	Final	1 Mar 2005 – 28 Feb 2011

4. TITLE AND SUBTITLE

Identification, Characterization and Clinical Development of the New Generation of Breast Cancer Susceptibility Alleles

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0204

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nazneen Rahman, M.D., Ph.D.

E-Mail: nazneen.rahman@icr.ac.uk

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Institute of Cancer Research
London SW7 3RP, United Kingdom

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Enter a brief (approximately 200 words) unclassified summary of the most significant finding during the research period. There is considerable evidence that genetic factors play an important role in causing breast cancer, but the genes involved in the majority of breast cancers are currently unknown. Our aim was to identify genetic factors that increase the risk of breast cancer occurring by performing analyses in our unparalleled series familial breast cancer samples. Using a candidate gene familial case-control design we identified three new breast cancer genes, ATM, PALB2 and BRIP1. By performing the largest genome-wide association analysis undertaken to date, in ~4000 familial breast cancer samples, we identified five new common genetic variants that predispose to breast cancer. In our final year we have optimized new sequencing technologies to analyse all genes (known as the 'exome') and undertook a pilot analysis of the exome in 20 familial breast cancer cases. We aim to use this technique in the future to uncover more of the genetic variants that cause breast cancer.

15. SUBJECT TERMS
Cancer genes, genetic predisposition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	68	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Breast cancer is a common disease in women but the causes are still largely unknown. There is considerable evidence to suggest that genetic factors play an important role in causing breast cancer. In the decade leading up to the start of this award considerable progress had been made and two major breast cancer genes, *BRCA1* and *BRCA2*, had been identified (reviewed in Walsh and King, 2007; Stratton and Rahman, 2008; Turnbull and Rahman, 2008 [1-3]). These genes carry a high risk of breast cancer (RR >10) but only account for a minority of breast cancer families and a very small proportion of breast cancer generally. Weaker genes were thought likely to be involved in the majority of familial breast cancers and some breast cancer cases without a family history of the disease, but few had been identified (Antoniou and Easton, 2003; Meijers-Heijboer et al. 2002) ([4, 5]).

The aim of my programme was to identify and characterize the genetic factors that increase the chance of breast cancer occurring. In order to achieve this I proposed analyses in an unparalled series of familial breast cancer cases that I have been collecting for the last decade. In the UK, Clinical Cancer Genetics is run through 26 Regional Genetic services and all genetic testing of breast cancer families is undertaken through this infrastructure. I have a study, known as the Familial Breast Cancer Study (FBCS), which recruits families with three or more cases of breast cancer through this clinical infrastructure. All of the Regional Genetic services participate in the study, and thus we have a high volume of referrals and national recruitment. DNA samples from over 5000 families, all curated with respect to clinical phenotype and *BRCA1/2* status, are available for our research. Using these unique sample resources we have successfully undertaken multiple different approaches, maximizing the extraordinary technological advancements that occurred over the course the programme to identify several new genetic variants that predispose to breast cancer, as detailed below.

Body

As part of the programme of work we defined five tasks. The outcome of these tasks is outlined in detail below.

Task 1: Evaluate the contribution of BRCA1 and BRCA2 exonic deletions and duplications to breast cancer susceptibility.

When we started the project it was unclear what proportion of *BRCA1* and *BRCA2* mutations were attributable to exonic deletions and duplications and the optimal method for their detection was also unclear. Such mutations are not typically detectable by standard PCR-based exonic amplification methods because the mutant allele is not amplified at all, and the sample therefore appears to be wild-type. Some form of copy number analysis is required. We undertook analysis for genomic exonic deletions and duplications of *BRCA1* and *BRCA2* in 1500 familial breast cancer cases from separate pedigrees in which small coding mutations of the genes had been excluded. We used a simple, cost-effective copy number analysis technique, multiplex ligation-dependent probe amplification (Schouten et al. 2002).

The analysis resulted in the identification of genomic duplication / deletion abnormalities in 4% of breast cancer families and demonstrated that:

- MLPA is a cheap, high-throughput and robust technique for copy-number variations, in most situations.
- MLPA should be undertaken in addition to sequencing in all breast cancer families.
- Certain probes showed inter-assay variability. We informed the manufacturers of this and problem and the probes were replaced.
- Single exon deletions must be further investigated and confirmed – firstly by sequencing to exclude a small exonic mutation under the probe, and if this is normal, by another copy-number assay such as quantitative PCR.

- The clinical features and risks of cancer are the same for families with genomic deletions / duplications as for intragenic mutations.

We subsequently changed our own *BRCA1/2* testing protocol so that we undertake MLPA analysis in addition to gene sequencing. This is also now the standard testing process used in clinical diagnostic testing laboratories throughout the UK and most of Europe.

Task 2. Perform familial case-control analyses of non-synonymous coding single nucleotide polymorphisms (SNPs) in DNA repair genes in familial breast cancer cases.

As part of the work that we undertook prior to starting the EOH programme we had sequenced DNA repair genes in 96 (1 tray) of index samples from *BRCA1/2* negative breast cancer families. This led to the identification of new breast cancer predisposition genes (see Task 5 below) and we also proposed to evaluate the 114 non-synonymous coding single nucleotide polymorphisms (SNPs) we identified in the DNA repair genes in larger case-control analyses to evaluate their contribution to breast cancer. When we started the project we anticipated that this would take several months with the extant technology. However, during the course of the programme there were substantial advancements in technology which allowed us to greatly improve our experimental design and to undertake much better powered, larger-scale experiments. Thus, instead of separately undertaking Task 2 and Task 4 we were able to undertake a single much larger experiment that included both the non-synonymous coding SNPs we had identified in our screen of DNA repair genes and all the other known non-synonymous coding SNPs. This experiment is described in detail in Task 4.

Task 3. Characterise the histopathology and immunohistochemistry of familial breast cancer.

This was one aspect of the programme where we were not able to achieve as much as I had hoped. There were substantial difficulties in acquiring the material to review and construct microarrays, in part because our local histopathology service underwent considerable

difficulties and changes during the programme. We therefore amended and tempered our aims to focus on the areas that we could achieve, that had maximum clinical utility, and that enhanced the other tasks. These are described below.

a) *Accruing data on the ER status of the cases included in the genome-wide association study so that analysis by ER status could be performed.*

As part of Task 4 we undertook a large genome-wide association study (see below). The early GWAS studies demonstrated that ER status is the strongest phenotypic surrogate (Easton et al. 2007). Therefore we prioritized obtaining hormonal receptor status on as many of the 4000 familial breast cancer cases included so that the analyses could be stratified by ER status. This demonstrated that for four of the SNPs (rs10995190, rs1011970, rs614367 and rs624797), the estimated per allele ORs were higher for ER-positive disease, with little association in ER-negative breast cancer, consistent with the pattern seen for the majority of breast cancer loci identified previously. For rs2380205 and rs704010, the per allele ORs for ER+positive and ER-negative disease were similar, but the number of ER-negative cases was too small to draw firm conclusions on the effect sizes for this subset (Turnbull et al. 2010; paper attached).

b) *Collection and analysis of sporadic breast cancer cases with triple-negative tumors (ER, PR and HER2 negative) which we are stratifying by BRCA1 status.*

Genetic and biological data indicate that triple-negative, basal-like tumors are a distinctive sub-phenotype of breast cancers that may have different underlying causes (Reis-Filho and Tutt, 2008). There is a known strong association of triple-negative tumor phenotype and *BRCA1* mutations (Atchley, et al. 2008). However, the contribution of *BRCA1* to triple-negative breast cancer in the absence of a strong family history remains unclear and is a source of considerable confusion diagnostically. In the final year of the grant we investigated this question by sequencing *BRCA1* in 308 individuals with triple-negative breast cancer. We

identified 45 *BRCA1* mutations in the 308 individuals (14.5%).This included 30 in the 149 selected series from genetics clinics or with young-age at onset (20.1%) and 15 /159 in the unselected series of cases from the breast cancer clinic (9.4%). There was strong age-effect with marked decrease in mutation frequency above 50 years in both the unselected and selected series.

	unselected	selected	
<30	0	9/29 (31%)	
30-39	5/22 (22%)	10/57 (17%)	
40-49	7/41 (17%)	7/32 (21%)	
50+	3/94 (3%)	4/31 (12%)	

These data suggest that the frequency of *BRCA1* mutations in individuals with TNT tumors diagnosed under 50 years is substantial and greater than the recommended threshold for testing (10% in most countries including US and UK). Thus it is appropriate to offer *BRCA1* testing to all women <50 years with TNT tumors. These data are also supportive with a recent paper suggesting that *BRCA* testing in TNT cases diagnosed under 50 years is cost-effective (Kwon et al, 2010). We are currently writing up these data and will submit for publication within the next three months.

c) *Tumor collection and pathological, immunohistochemical and loss of heterozygosity analyses to define the tumor characteristics associated with the rare, intermediate-penetrance breast cancer susceptibility genes, ATM, BRIP1, CHEK2, PALB2.*

We are still committed to this project, which we believe will be interesting. To date we have collected tumor material from only 20 cases, though we have pathology reports from many additional cases. There has been considerable activity in analyzing these genes around the world since our gene discovery papers and therefore we are planning to engage in international collaborative initiatives to take this project forward in the future.

Task 4. Perform genome-wide familial case-control analyses of non-synonymous coding SNPs,

As described above, we altered the design of our study to take advantage of technological advancements and substantial decreases in cost. In collaboration with the Wellcome Trust Case Control Consortium (WTCCC) we analyzed 14,471 non-synonymous SNPs in 864 familial *BRCA1/2*-negative cases and 1498 controls, using a custom array. This array included all the known non-synonymous SNPs in the databases that it was possible to design probes for at that time, together with the ns-SNPs we had discovered through our DNA repair mutational analyses. This analysis did not reveal any breast cancer predisposition alleles, though the overall experiment did identify variants associated with auto-immune diseases (WTCCC, 2007). Concurrently, while we were undertaking this experiment the first genome-wide association study (GWAS) in breast cancer (in which we collaborated supplying ~half of the samples in the first stage) demonstrated that common variants associated with breast cancer were typically NOT coding variants (Easton et al. 2007). We therefore altered our strategy to focus on undertaking a larger GWAS to look for common variants throughout the genome rather than focusing on coding variants. These experiments are outlined under Task 5.

Task 5. Identify and characterize lower-penetrance breast cancer susceptibility alleles

a) *Undertake case-control resequencing of genes to identify further rare, intermediate/low-penetrance genes.*

The initial aim of our breast cancer work was to extend the familial case-control approach that we had successfully utilized to identify *CHEK2* as an intermediate breast cancer predisposition gene (RR 2-3) to identify further DNA repair genes that predispose to breast cancer (Miejers-Heijboer, 2002; The CHEK2 breast cancer consortium). There was considerable epidemiological evidence to suggest that mutations in *ATM* might contribute to breast cancer, but the molecular proof had been lacking. Therefore, in the first instance, we

undertook a familial breast cancer case-control analysis to demonstrate that inactivating mutations in *ATM* are intermediate breast cancer predisposition gene and to clarify an issue that had been highly controversial for nearly 20 years (Renwick et al, 2006, paper attached; Ahmed and Rahman, 2006).

We also selected further DNA repair genes with close functional links to BRCA1 and or BRCA2 for analysis in a familial case-control analysis and demonstrated that *BRIP1* (also known as *BACH1*) is an intermediate breast cancer predisposition gene (Seal et al. 2006, paper attached). Shortly thereafter a new gene *PALB2*, that encodes a protein that interacts with BRCA2, was identified (Xia et al. 2006). We demonstrated *PALB2* mutations predispose to breast cancer and independently this was reported in Finnish breast cancer cases (Rahman et al, 2007 paper attached; Errko et al. 2007). Separately, through my funding for childhood cancer genetics I demonstrated that biallelic *PALB2* mutations cause a severe form of Fanconi anemia, similar to that that seen in biallelic *BRCA2* mutation carriers (Reid et al. 2007). We later, also evaluated a new DNA repair gene, *GEN1*, which was identified as a key Holliday junction resolvase involved in homologous recombination that had been proposed to be a breast cancer predisposition gene (Ip et al. 2008). Our data showed that despite its role in DNA repair *GEN1* variants do not act as susceptibility alleles analogous to *CHEK2*, *ATM*, *BRIP1*, and *PALB2* (Turnbull et al. 2010 paper attached).

b) *Undertake a second-generation genome-wide association study to identify common, low-penetrance breast cancer susceptibility alleles.*

In 2007, we collaborated with Professor Douglas Easton to complete the first GWAS in breast cancer. This utilized 400 genetically enriched breast cancer cases and 400 controls typed for over 220,000 SNPs. These SNPs were correlated with ~71% of known common SNPs, at r^2>0.5. Putative associations were followed up in ~26,000 cases and 26,000 controls. This study provided clear evidence for five novel breast cancer susceptibility loci (Easton et al.

2007). Further studies led to the identification of a further 5 loci and together these 10 loci explained about 6% of the familial risk of the disease (reviewed in Turnbull and Rahman, 2008).

Although the GWAS studies undertaken had been successful they were underpowered as the genome-wide phase was undertaken in relatively small series. We successfully applied to Wellcome Trust to obtain funding to support the genotyping costs of a second-generation GWAS scan. We genotyped 582,886 SNPs in 3,659 cases enriched for a family history of the disease and compared the data to genotypes from 4,897 controls. We evaluated promising associations in a second Stage in a collaborative analysis comprising 12,576 cases and 12,223 controls. We identified five novel susceptibility loci, on chromosomes 6, 10 and 11 (P=3.7x10^{-7} to P=4.6x10^{-16}). We also identified SNPs in the *ESR1,* 8q24 and *LSP1* that were more strongly associated with risk than those reported previously. Known susceptibility loci exhibited stronger associations in our study than in population-based studies, consistent with polygenic susceptibility to the disease, and confirming that our strategy of using familial cases enhances the power of gene discovery experiments (Turnbull et al. 2010, paper attached).

Based on the estimated per allele ORs from stage 2 of our study, the newly identified loci explain approximately 1.2% of the familial risk of breast cancer, though the overall contribution may be larger, since the true causal variants may be more strongly associated with disease than the SNPs tagging them. Taken together with estimates from previous studies, the 18 confirmed breast cancer susceptibility loci together explain approximately 8% of the familial risk of breast cancer, while rarer mutations in the known high risk (principally *BRCA1* and *BRCA2*) and moderate risk loci explain a further ~20%. This was, by far, the largest breast cancer GWAS to date and confirms that the *FGFR2* and *TOX3* loci (conferring per allele ORs 1.2-1.3) are the strongest common susceptibility loci that are detectable with high coverage genome-wide tagSNP sets. The residual familial risk is therefore likely to be due to a combination of a large number of common variants with smaller

effects, together with rarer variants not testable with current arrays, but potentially identifiable through sequencing strategies.

c) *Undertake a study to evaluate the contribution of common CNVs to breast cancer predisposition*

There has been considerable interest in the contribution of polymorphic copy number variants (CNVs) to disease susceptibility . In collaboration with the WTCCC we undertook a large experiment evaluating 3,432 polymorphic CNVs, including an estimated 50% of all common CNVs larger than 500bp, in 2000 of our familial breast cancer cases. The CNV array analyses were outsourced and funded by the WTCCC. There were two CNVS of potential interest in breast cancer and we evaluated these in-house using real-time PCR supported by the EOH programme. Both were shown to be false-positives and overall the experiment demonstrated that common CNVs are unlikely to contribute substantially to breast cancer (or indeed any of the other phenotypes included in the experiment). (paper attached, WTCCC, 2009).

d) *Undertake exomic analyses to identify breast cancer predisposition genes*

In our original application we aimed to undertake candidate gene case-control resequencing of genes, focusing on DNA repair genes. The original strategy involved sequencing 96 (1 tray) familial breast cancer cases through the full gene and undertaking additional sequencing of genes in which we identified truncating variants in larger series of cases and controls (typically 1000 cases and 1000 controls). This strategy led to the identification of three new rare, low-intermediate penetrance genes, as outlined above. However, analysis of further DNA repair genes did not lead to the identification of additional genes. The availability of new sequencing technologies together with pulldown arrays targeting the exons of all genes (known as the 'exome') has made it feasible to progress from a candidate to a genome-wide gene resequencing strategy. Within the final year of the grant we have been engaged in optimizing the new sequencing technologies. We conducted a pilot analysis of exomes in 20 familial

breast cancer cases, including some cases with mutations in known genes, which were readily detectable (Snape et al, 2010). The pilot experiment has allowed us to optimize all of the laboratory, IT and analytical infrastructure and we are now well-positioned to extend these analyses and undertake much larger-scale experiments. We are seeking funding for this currently and these endeavours will form the crux of our future work to identify breast cancer predisposition genes.

Key Research Accomplishments

- Seven publications in high-ranking journals, including five in Nature Genetics and one in Nature.
- Identification of three intermediate-penetrance breast cancer predisposition genes, *ATM*, *BRIP1* and *PALB2*, through familial case-control resequencing experiments.
- Identification of five common variants that are low-penetrance breast cancer predisposition alleles through the largest genome-wide association study performed in breast cancer to date.
- Successful application for funding for the genotyping required for the genome-wide association study.
- Training of four Clinical Research Fellows in Cancer Genetics

Reportable Outcomes

Published Manuscripts

1. Renwick, A, Thompson, D, Seal, S, Kelly, P, Chagtai, T, Ahmed, M, North, B, Jayatilake, H, Barfoot, R, Spanova, K, McGuffog, L, Evans, D G, Eccles, D, Easton, D F, Stratton, M R, and Rahman, N. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature Genetics* 38:873-5

2. Seal, S, Thompson, D, Renwick, A, Elliott, A, Kelly, P, Barfoot, R, Chagtai, T, Jayatilake, H, Ahmed, M, Spanova, K, North, B, McGuffog, L, Evans, D G, Eccles, D, Easton, D F, Stratton, M R, and Rahman, N. (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nature Genetics* 38:1239-41.

3. Wellcome Trust Case Control Consortium and The Australo-Angle-American Spondylitis Consortium Association (2007) Scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants *Nature Genetics* 39:1329-1338

4. Rahman, N, Seal, S, Thompson, D, Kelly, P, Renwick, A, Elliott, A, Reid, S, Spanova, K, Barfoot, R, Chagtai, T, Jayatilake, H, McGuffog, L, Hanks, S, Evans, D G, Eccles, D, Easton, D F, and Stratton, M R. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature Genetics* 39:165-7.

5. Turnbull C, Hines S, Renwick A, Hughes D, Pernet D, Elliott A, Seal S, Warren-Perry M, Evans DG, Eccles D, Breast Cancer Susceptibility Collaboration (K), Stratton MR and Rahman N (2010) Mutation and association analysis of *GEN1* in breast cancer susceptibility. *Breast Cancer Research and Treatment* 124:283-8

6. Wellcome Trust Case Control Consortium. (2010) Genome-wide association studies of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464: 713-720.

7. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS, Hughes D, Warren-Perry M, Tapper W, Eccles D, Evans DG, The Breast Cancer Susceptibility Collaboration UK, Hooning M, Schutte M, ven del Ouweland A, Houlston R, Ross G, Langford C, Pharoah PDP, Stratton MR, Dunning AM, Rahman N, Easton DF (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics* 42:504-507

Presentations

2006		
Goulstonian Prize Lecture, Royal College of Physicians		Finding Cancer predisposition genes – past lessons and future challenges
British Society of Human Genetics		Low penetrance breast cancer genes – Plenary Session
National Cancer Research Institute	Invited lecture	DNA repair and breast cancer susceptibility – a complex web of high and low penetrance alleles

Fanconi Anemia Research Fund annual conference	Invited lecture	PALB2 mutations cause Fanconi anemia FA-N and predispose to childhood cancer
2007		
London IDEAS conference, Institute of Child Health	Invited lecture	Rare, intermediate penetrance breast cancer susceptibility genes
Ovarian Cancer Association Collaboration Annual Meeting	Invited lecture	Rare, intermediate penetrance breast cancer genes
Clare Hall	Invited lecture	New links between DNA repair genes and cancer susceptibility
CNIO, Madrid	Invited lecture	Breast Cancer Susceptibility Genes
2008		
Breast Cancer Association Consortium meeting, Barcelona	Invited lecture	Rare, intermediate penetrance breast cancer genes
IARC-EACR-AACR Integrative Molecular Cancer Epidemiology symposium	Invited lecture	Case-control mutation screening to identify breast cancer predisposition genes
2009		
Genetics Society, London	Invited lecture	Clinical utility of breast cancer genes: current practice and future prospects
Institute of Cancer Research Centenary Conference	Invited lecture	Cancer Predisposition Genes - from cause to clinic
2010		
American Association for Cancer Research, 101st Annual Meeting - Washington	Invited speaker	Genetic Predisposition Breast Cancer - New Discoveries and their Clinical Applications
Wellcome Trust: 10 Years of Genomic Medicine	Invited Speaker	BRCA1 and BRCA2 - from gene to the creation of a clinical specialty
ICR / Royal Marsden Annual Research Report Launch	Invited Speaker	Predicting cancer: where we are and where we need to get to
2011		

2011 4th Annual Royal Marsden Breast Cancer Meeting	Invited Speaker	Identifying at risk patients for genetic testing
Wellcome Trust Biomedicine Forum	Invited Speaker	Genetics and the NHS

Degrees Awarded

Four clinical research fellows have been supported by this award:

Clare Turnbull is an exceptional clinician-scientist. She started as clinical fellow supported by the EOH programme and was then able to obtain a highly prestigious personal training fellowship which has supported her salary. She will shortly submit her PhD and is planning to become an academic clinician and has an extremely bright future.

Munaza Ahmed has been awarded her higher degree. She successfully applied for a clinical training position in genetics after undertaking research with me. She has now completed this and has recently been appointed to a substantive position as a Cancer Geneticist, primarily managing women with a family history of breast cancer.

Helen Hanson will have the viva for her higher degree shortly. She is a gifted clinical academic and is particularly engaged in clinical translation of risk estimation. We are planning to appoint to run Cancer Genetic clinics once she has completed her clinical training next year.

Lisa Robertson spent a year with us and undertook the triple-negative breast cancer study which she is currently writing-up.

Individuals supported by the award

Name	Period
Statisticians/database manager/non-lab support	
B North	06/2005 – 12/2005
Anna Elliott	11/2005 – 09/2007
Ann Strydom	02/2009 – 02/2010
Fiona Harvey	01/2008 – 11/2008
F M G Pearl	11/2007 – 06/2008
Richard Bowman	01/2009 – 10/2009
Research Assistants	
Darshna Dudakia	06/2008 – 01/2010
S R Meka	08/2007 – 01/2008

Anne Murray	08/2007 – 06/2008
Deborah Hughes	10/2007 – 05/2010
Karen Barker	07/2005 – 06/2008
Clinical Research Fellows	
Muna Ahmed	08/2005 – 12/2007
Clare Turnbull	01/2007 – 10/2007
Helen Hanson	10/2008 – 02/2011
Lisa Robertson	11/2009 – 03/2010 & 07/2010 – 09/2010

Funding successfully applied for based on work supported by this award

I successfully applied to the Wellco me Trust to fund our second generation genome-wide association study and was awarded £655,500.

Conclusion

Through the EOH programme I have been able to establish my group as one of the world leaders in the identification and characterization of breast cancer predisposition genes. Through the course of the programme we were able to make substantial contributions to the area, delineating the genetic landscape of breast cancer into three strata; rare high-penetrance genes; rare, intermediate-penetrance genes and common, low-penetrance genes. Together the known genes contribute ~70% of the genetic risk of breast cancer, and thus there is still much work to be done, to identifying new genes and to translating the findings to clinical benefit. New technological advances will allow us to undertake large-scale sequencing initiatives towards this aim and we are hopeful of further successes in the future.

The discovery of the genetic causes of breast cancer are important for diagnosis and management of women affected with breast cancer and also can give important information for unaffected women, assisting in identifying those women at increased risk, who may benefit from surveillance and preventative strategies.

References

Ahmed M and Rahman N (2006) ATM and breast cancer susceptibility Oncogene reviews 25:5906-5911

Antoniou AC and Easton DF (2003) Polygenic inheritance of breast cancer: Implications for design of association studies. Genet. Epidemiol. 25:190-202.

Atchley DP et al. (2008) Clinical and pathologic characteristics of patients with BRCA-positive and BRCA-negative breast cancer J Clin Oncol. 26: 4282-4288

The CHEK2 breast cancer consortium (2004) CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet. 74:1175-1182

Easton DF et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-1093

Erkko H et al (2007) A recurrent mutation in PALB2 in Finnish cancer families. Nature. 446: 316-319.

Ip SC et al (2008). Identification of Holliday junction resolvases from humans and yeast. Nature, 456: 357-361.

Kwon JS et al (2010) Expanding the criteria for BRCA mutation testing in breast cancer survivors. J Clin Oncol.28:4214-4120

Meijers-Heijboer H, et al (2002) Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nature Genetics 31:55-59.

Rahman N et al (2007). PALB2, which encodes a BRCA2 interacting protein, is a breast cancer susceptibility gene. Nature Genetics 39:165-167

Reid S et al (2007 Biallelic mutations in PALB2, which encodes a BRCA2 interacting protein, cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nature Genetics 39:162-164

Reis-Filho, J.S and Tutt A (2008) Triple negative tumours: a critical review. Histopathology 52(: 108-118.

Renwick A et al (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 38:873-875

Schouten JP et al (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 30:e57.

Seal S, et al (2006) Truncating mutations in BRIP1 are low penetrance breast cancer susceptibility alleles. Nature Genetics 38:1239-1241

Snape, K., et al (2010) Exome sequencing in the identification of breast cancer predisposition genes, in 60th Annual American Society of Human Genetics Meeting. Washington.

Stratton MR and Rahman N (2008) The emerging landscape of breast cancer susceptibility Nature Genetics 40:17-22

Turnbull, C. and N. Rahman (2008) Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet,. 9:321-345.

Turnbull, C. et al (2010). Mutation and association analysis of *GEN1* in breast cancer susceptibility. Breast Cancer Res Treat,.

Turnbull, C., et al. (2010), Genome-wide association study identifies five new breast cancer susceptibility loci. Nature Genetics 42: p. 504-507.

Walsh, T. and M.C. King (2007) Ten genes for inherited breast cancer. Cancer Cell 11:103-105.

Wellcome Trust Case Control Consortium and The Australo-Angle-American Spondylitis Consortium Association (2007) Scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants Nature Genetics 39:1329-1338

Wellcome Trust Case Control, Consortium (2010), Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464:713-720.

Xia, B., et al. (2006) Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. Mol.Cell, 22:719-729

Appendices

Seven manuscripts resulting from work undertaken through this programme and one review of breast cancer genetics are attached.

nature genetics

ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles

Anthony Renwick[1], Deborah Thompson[2], Sheila Seal[1], Patrick Kelly[1], Tasnim Chagtai[1], Munaza Ahmed[1], Bernard North[1], Hiran Jayatilake[1], Rita Barfoot[1], Katarina Spanova[1], Lesley McGuffog[2], D Gareth Evans[3], Diana Eccles[4], The Breast Cancer Susceptibility Collaboration (UK), Douglas F Easton[2], Michael R Stratton[1,5] & Nazneen Rahman[1]

We screened individuals from 443 familial breast cancer pedigrees and 521 controls for *ATM* sequence variants and identified 12 mutations in affected individuals and two in controls ($P = 0.0047$). The results demonstrate that *ATM* mutations that cause ataxia-telangiectasia in biallelic carriers are breast cancer susceptibility alleles in monoallelic carriers, with an estimated relative risk of 2.37 (95% confidence interval (c.i.) = 1.51 3.78, $P = 0.0003$). There was no evidence that other classes of *ATM* variant confer a risk of breast cancer.

ATM is a protein kinase that has a key role in monitoring and repair of double strand DNA breaks. Biallelic mutations in *ATM* cause the autosomal recessive disease ataxia telangiectasia. Over 70% of *ATM* mutations that cause ataxia telangiectasia are base substitutions, insertions or deletions that generate premature termination codons or splicing abnormalities[1] (see http://www.benaroyaresearch.org/bri investigators/atm.htm). Studies of individuals with ataxia telangiectasia have suggested that female relatives heterozygous for an *ATM* mutation have a two to fivefold increase in risk of breast cancer[2,3]. A key prediction of this hypothesis is that heterozygosity for *ATM* mutations (that is, heterozygosity for variants in *ATM* that cause ataxia telangiectasia) is more common among individuals with breast cancer than the general population. However, studies of breast cancer case and control series have failed to show an elevated frequency of truncating *ATM* mutations in individuals with breast cancer[4–6]. These results have prompted alternative models of the role of *ATM* in breast cancer susceptibility. It has been proposed that missense variants (in particular, variants that do not cause ataxia telangiectasia) predispose to breast cancer[7]. It has also been suggested that only a subset of *ATM* mutations, defined by specific biological characteristics, confer a risk of breast cancer, and that this risk is high, similar to that of mutations in *BRCA1* and *BRCA2* (ref. 8). Finally, it has been proposed that the elevated frequency of breast cancer in mothers of individuals with ataxia telangiectasia is related to factors other than heterozygosity for *ATM* mutations[9].

To resolve the confusion regarding the role of *ATM* mutations in breast cancer susceptibility, we adopted a case control strategy. To maximize the power of the study, we incorporated the following design features. First, we screened genomic DNA from all cases and controls for mutations through the 62 coding exons and splice junctions of *ATM* (**Supplementary Methods** and **Supplementary Table 1** online). This allowed direct and unbiased comparison of the mutation frequency and spectrum in cases and controls. Second, we included only index cases from families with at least three breast cancers. The use of familial, rather than sporadic, breast cancers cases increases the power substantially, as previously illustrated in studies of

Table 1 *ATM* mutations identified in familial breast cancer cases and controls

Family	Mutation	Effect	Number of cases ($n = 443$)	Number of controls ($n = 521$)
1	8264delATAAG (8152del117)	Exon 58 skipped	1	0
2	IVS40 1050A→G (5762ins137)	Premature truncation	1	0
3	IVS44+1G→A (6096del103)	Premature truncation	1	0
4	3802delG	Premature truncation	1	0
5	C3349T	Q1117X	1	0
6	5290delC	Premature truncation	1	0
7	790delT	Premature truncation	1	0
8	C7311A	Y2437X	1	0
9	IVS59+1delGTGA (8269del150)	Exon 59 skipped	1	0
10, 11	T7271G	V2424G	2	0
12	TG8565 8566AA	SV2855 2856RI	1	0
	C802T	Q268X	0	1
	6997insA	Premature truncation	0	1

The mutations identified in families 1, 2, 3, 4, 6, 7, 9 10, 11 and 12 have previously been reported as causative in ataxia-telangiectasia cases[3,8,13] (http://www.benaroyaresearch.org/bri_investigators/atm.htm). The effect on the transcript of mutations in families 1, 2, 3 and 9 have previously been investigated by RT-PCR and sequencing and are annotated in parentheses after the mutation. The pedigrees of families 1–12 are shown in **Figure 1**.

[1]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, UK. [2]Cancer Research UK, Genetic Epidemiology Unit, Strangeways Research Laboratories, University of Cambridge, CB1 8RN, UK. [3]Department of Medical Genetics, St Mary's Hospital, Manchester, M13 0JH, UK. [4]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, SO16 6YA, UK. [5]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. Correspondence should be addressed to N.R. (nazneen.rahman@icr.ac.uk).

Received 27 March; accepted 9 June; published online 9 July 2006; doi:10.1038/ng1837

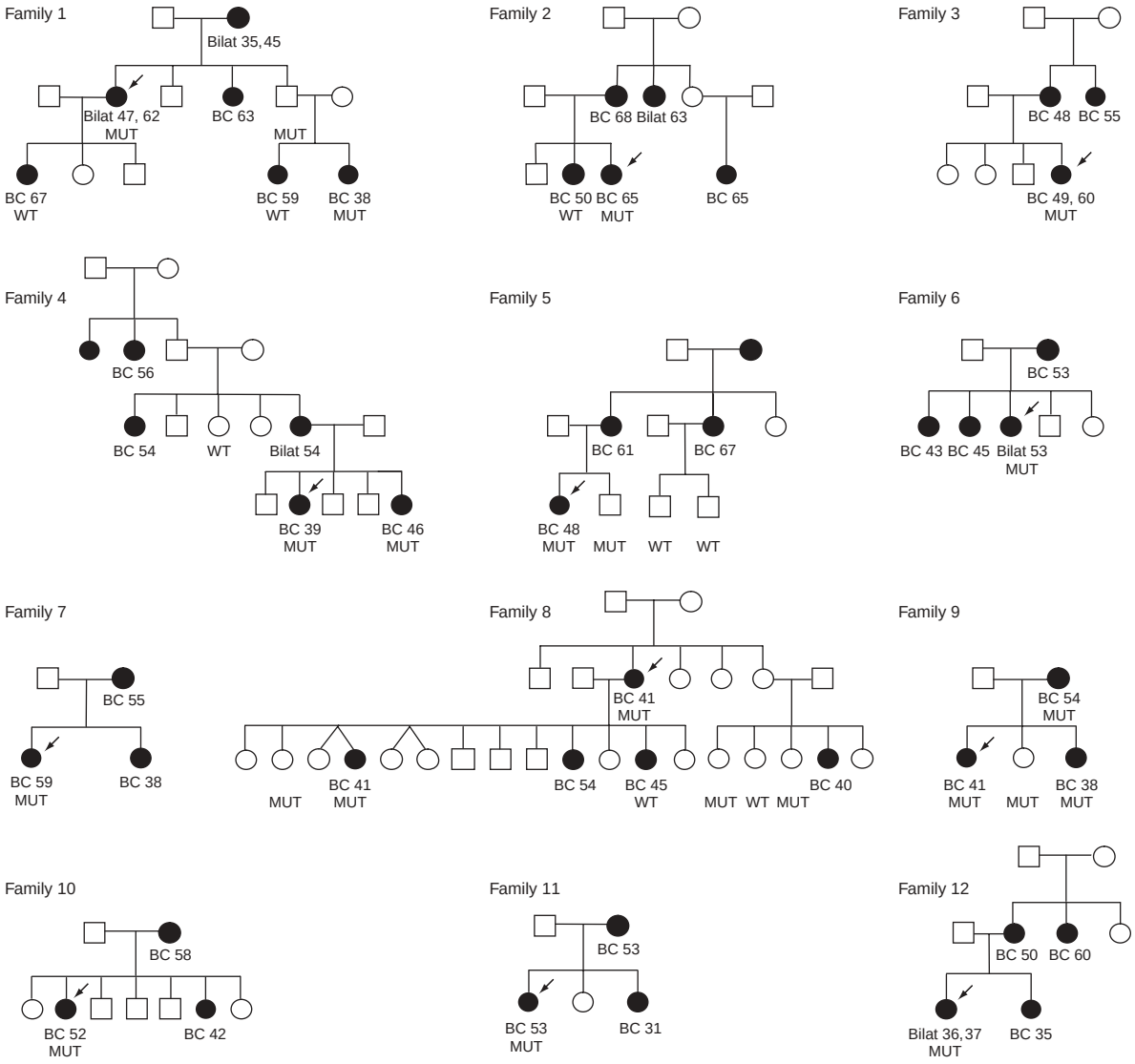


Figure 1 Abridged pedigrees of twelve breast cancer families with *ATM* mutations. Individuals with breast cancer are shown as filled circles, with the age at diagnosis given underneath. If the individual had metachronous bilateral breast cancer, two ages are given. Other cancers or medical conditions are not shown. The index case that was initially screened through *ATM* is shown by an arrow. The *ATM* mutation in each family is given in **Table 1**. BC, breast cancer; Bilat, bilateral breast cancer; MUT, *ATM* mutation present; WT, *ATM* mutation absent.

CHEK2 mutations in breast cancer[10–12]. Finally, the familial case series had already been pre screened for *BRCA1* and *BRCA2* mutations and large deletions and duplications. Familial cases due to *BRCA1* or *BRCA2* were excluded, thus enriching the case series for other breast cancer susceptibility alleles (**Supplementary Methods**).

We identified nine (2.04%) *ATM* mutations that result in premature truncation or exon skipping in 443 familial breast cancer cases and two truncating mutations (0.4%) in 521 controls ($P = 0.028$; **Table 1** and **Fig. 1**). All of the mutations are predicted to cause ataxia telangiecta sia, and seven of the nine mutations identified in cases have previously been reported in ataxia telangiectasia families, including the two most common mutations in the UK, 5762ins137 and 3802delG. The frequency of heterozygotes for truncating *ATM* mutations observed in the control series (0.5%, allowing for a mutation screening sensitivity of 70%) is consistent with that previously estimated for the UK population based on the incidence of ataxia telangiectasia[3].

We also identified 37 different missense variants (**Supplementary Table 2** online). There is strong prior evidence that two of these, V2424G and SV2855 2856RI, are pathogenic mutations in individuals with ataxia telangiectasia[3,8,13] (see also http://www.benaroyaresearch. org/bri investigators/atm htm and **Supplementary Note** online). Excluding V2424G and SV2855 2856RI, we identified 35 nonsynony mous missense variants, of which 12 were present in both cases and controls, 13 were present exclusively in cases and 10 were present exclusively in controls. None of these has previously been implicated as a disease causing ataxia telangiectasia mutation. Five variants (S49C, F858L, P1054R, L1420F, D1853N) had a minor allele frequency of >1% in the combined set; the difference in carrier frequencies between cases and controls was not statistically significant for any of these. Of the remaining 30 rare nonsynonymous missense variants, we found 26 instances in 25 cases, compared with 21 instances in 19 controls ($P = 0.16$). Furthermore, there was no evidence of clustering

of rare nonsynonymous missense variants within conserved ATM functional domains or in the predicted pathogenicity of the variants in cases compared with controls (**Supplementary Note**).

Combining *ATM* truncating, splicing and missense mutations for which there is strong prior evidence of involvement in ataxia telangiectasia, there were 12 mutations in cases and two in controls ($P = 0.0047$; **Table 1**). The relative risk of breast cancer associated with *ATM* mutations was estimated to be 2.37 (95% c.i. = 1.51 3.78, $P = 0.0003$) by segregation analysis incorporating information from the controls and the full pedigrees of the cases (**Supplementary Methods** and **Supplementary Note**). This estimate is consistent with those derived from studies of ataxia telangiectasia families and is equivalent to a breast cancer population attributable fraction of 0.86% (95% c.i. = 0.32% 1.72%). There was no evidence of a difference in relative risk between carriers aged below or above 50 years ($P = 0.74$), although the estimated relative risk below age 50 (2.50, 95% c.i. = 1.41 4.17) is consistent with the more substantial risks at young ages suggested by some studies of ataxia telangiectasia families[3]. Consistent with the modest estimated relative risk, there was limited evidence of cosegregation of breast cancer with the *ATM* mutation in the five families from which additional samples were available, with five of the nine tested additional individuals with breast cancer carrying the *ATM* mutation present in that family (four expected if the *ATM* mutation were unrelated to breast cancer, $P = 0.36$; **Fig. 1**).

We compared the extent of breast cancer clustering, age at diagnosis and frequency of bilateral breast cancer in index cases with and without *ATM* mutations. The family history of breast cancer was slightly, but not significantly, higher in individuals with *ATM* mutations (median family history score 2.75 versus 2.25, $P = 0.21$). There was no difference in the median age at diagnosis of index cases with an *ATM* mutation (48.6 years) compared with index cases without an *ATM* mutation (48.9 years). The frequency of bilateral cancers was also similar: 1 out of 12 (8%) index cases with an *ATM* mutation developed metachronous bilateral breast cancer, compared with 49/431 (11%) index cases without an *ATM* mutation.

We have previously demonstrated that a truncating mutation in *CHEK2* (*CHEK2**1100delC) is a breast cancer susceptibility allele conferring a twofold relative risk[10,11]. We screened the 443 cases and 521 controls in this study for *CHEK2**1100delC and identified 13 cases and three controls with the mutation ($P = 0.0048$). None of the *ATM* mutation carriers also carried *CHEK2**1100delC. These data indicate that, in the UK population, the combined *ATM* mutation prevalence is similar to that of *CHEK2**1100delC; both are associated with similar risks of breast cancer; and both make a similar contribution to breast cancer incidence.

The role of *ATM* in breast cancer susceptibility has been controversial for nearly 20 years. We have now provided strong evidence that *ATM* mutations that cause ataxia telangiectasia are breast cancer susceptibility alleles. This result is fully consistent with studies of ataxia telangiectasia families. We did not find evidence of a risk associated with sequence variants not predicted to cause ataxia telangiectasia. Although we cannot rule out some variation in risk by mutation, the data are consistent with an approximately twofold increase in risk of breast cancer associated with all *ATM* mutations that cause ataxia telangiectasia.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the participating families who were recruited to the study by the Familial Breast Cancer Collaboration (UK), which includes the following contributors: A. Ardern-Jones, J. Berg, A. Brady, N. Bradshaw, C. Brewer, G. Brice, B. Bullman, J. Campbell, B. Castle, R. Cetnarsryj, C. Chapman, C. Chu, N. Coates, T. Cole, R. Davidson, A. Donaldson, H. Dorkins, F. Douglas, D. Eccles, R. Eeles, F. Elmslie, D.G. Evans, S. Goff, S. Goodman, D. Goudie, J. Gray, L. Greenhalgh, H. Gregory, N. Haites, S.V. Hodgson, T. Homfray, R.S. Houlston, L. Izatt, L. Jeffers, V. Johnson-Roffey, F. Kavalier, C. Kirk, F. Lalloo, I. Locke, M. Longmuir, J. Mackay, A. Magee, S. Mansour, Z. Miedzybrodzka, J. Miller, P. Morrison, V. Murday, J. Paterson, M. Porteous, N. Rahman, M. Rogers, S. Rowe, S. Shanley, A. Saggar, G. Scott, L. Side, L. Snadden, M. Steel, M. Thomas, S. Thomas. We thank A. Hall and E. Mackie for coordination of sample collection. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. M. Ahmed and B. North are supported by the US Army Medical Research and Material Command grant #W81XWH-05-1-0204. This research was supported by the Institute of Cancer Research and by grants from the Breast Cancer Campaign and Cancer Research UK.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Shiloh, Y. *Nat. Rev. Cancer* **3**, 155 168 (2003).
2. Swift, M. *et al. N. Engl. J. Med.* **316**, 1289 1294 (1987).
3. Thompson, D. *et al. J. Natl. Cancer Inst.* **97**, 813 822 (2005).
4. Fitzgerald, M.G. *et al. Nat. Genet.* **15**, 307 310 (1997).
5. Teraoka, S.N. *et al. Cancer* **92**, 479 487 (2001).
6. Sommer, S.S. *et al. Cancer Genet. Cytogenet.* **145**, 115 120 (2003).
7. Gatti, R.A. *et al. Mol. Genet. Metab.* **68**, 419 423 (1999).
8. Chenevix-Trench, G. *et al. J. Natl. Cancer Inst.* **94**, 205 215 (2002).
9. Olsen, R.H. *et al. Br. J. Cancer* **93**, 260 265 (2005).
10. Meijers-Heijboer, H. *et al. Nat. Genet.* **31**, 55 59 (2002).
11. CHEK2 breast cancer case-control consortium. *Am. J. Hum. Genet.* **74**, 1175 1182 (2004).
12. Antoniou, A.C. *et al. Genet. Epidemiol.* **25**, 190 202 (2003).
13. Becker-Catania, S.G. *et al. Mol. Genet. Metab.* **70**, 122 133 (2000).

nature genetics

Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles

Sheila Seal[1], Deborah Thompson[2], Anthony Renwick[1], Anna Elliott[1], Patrick Kelly[1], Rita Barfoot[1], Tasnim Chagtai[1], Hiran Jayatilake[1], Munaza Ahmed[1], Katarina Spanova[1], Bernard North[1], Lesley McGuffog[2], D Gareth Evans[3], Diana Eccles[4], The Breast Cancer Susceptibility Collaboration (UK), Douglas F Easton[2], Michael R Stratton[1,5] & Nazneen Rahman[1]

We identified constitutional truncating mutations of the BRCA1-interacting helicase *BRIP1* in 9/1,212 individuals with breast cancer from *BRCA1*/*BRCA2* mutation negative families but in only 2/2,081 controls ($P = 0.0030$), and we estimate that *BRIP1* mutations confer a relative risk of breast cancer of 2.0 (95% confidence interval = 1.2 3.2, $P = 0.012$). Biallelic *BRIP1* mutations were recently shown to cause Fanconi anemia complementation group J. Thus, inactivating truncating mutations of *BRIP1*, similar to those in *BRCA2*, cause Fanconi anemia in biallelic carriers and confer susceptibility to breast cancer in monoallelic carriers.

Breast cancer is approximately twice as common in sisters and mothers of affected individuals as in the general population. Inacti vating mutations in *BRCA1*, *BRCA2*, and *TP53* confer a high risk of developing breast cancer (10 to 20 fold by age 60), whereas inactivat ing mutations of *CHEK2* and *ATM* are associated with more modest risks (approximately twofold). Together, these susceptibility genes are estimated to account for up to 25% of the familial risk of breast cancer. Therefore, most familial aggregation of breast cancer remains unexplained[1].

To identify additional breast cancer susceptibility genes, we screened several genes encoding proteins that interact with the products of known breast cancer predisposition genes. *BRIP1* (also known as *BACH1*) encodes a DEAH helicase that interacts with the BRCT domain of BRCA1 and has BRCA1 dependent DNA repair and checkpoint functions[2,3]. Inactivating mutations in *BRCA1* predispose to breast cancer. Inactivation of BRIP1 results in abrogation of certain BRCA1 functions, and therefore it is plausible that inactivating *BRIP1* mutations also predispose to breast cancer[4,5]. To investigate this hypothesis, we screened the full coding sequence and intron exon boundaries of *BRIP1* by conformation sensitive gel electrophoresis (CSGE) in genomic DNA from 1,212 women with breast cancer and 2,081 controls (**Supplementary Methods** and **Supplementary Table 1** online). All the individuals with breast cancer had a family history of at least one first degree relative with breast cancer or equivalent and/or a relative with ovarian cancer. Additionally, all affected individuals were negative for mutations and large deletions or duplications of *BRCA1* and *BRCA2* (see **Supplementary Methods** for full description of case and control series and mutational analyses of *BRCA1*, *BRCA2* and *BRIP1*). The use of this familial case control design increases the power substantially[1].

We identified five different truncating mutations in nine of the 1,212 individuals with breast cancer, compared with two truncating mutations in the 2,081 controls ($P = 0.0030$; **Table 1** and **Fig. 1**). There was no evidence of a difference in likelihood of carrying a *BRIP1* mutation between probands with bilateral or unilateral cancers ($P = 0.63$) or by extent of family history of breast cancer ($P = 0.31$). We estimated the relative risk of breast cancer associated with truncating *BRIP1* mutations to be 2.0 (95% confidence interval (c.i.) = 1.2 3.2; $P = 0.012$) by segregation analysis, incorporating information from the controls and the full pedigrees of the affected individuals (**Supplementary Methods**). The relative risk for carriers aged less than 50 years was 3.5 (95% c.i. = 1.9 5.7), which was significantly higher than the relative risk for carriers above this age ($P = 0.020$). Consistent with the modest estimated relative risk,

Table 1 *BRIP1* mutations identified in individuals with breast cancer and controls

Family	Mutation	Effect	Number of affected individuals ($n = 1,212$)	Number of controls ($n = 2,081$)
1	141delC	Premature truncation	1	0
2 6	2392C→T	R798X	5	1
7	IVS17+2insT	Exon 17 or exon 18 skipped	1	0
8	2008insT	Premature truncation	1	0
9	2255delAA	Premature truncation	1	0
	2108delAinsTCC	Premature truncation	0	1

The mutations identified in families 2 6, 7 and 9 have previously been reported as causative in Fanconi anemia subtype J[8 10]. The effect on the transcript of the mutation in family 3 has previously been investigated by RT-PCR and sequencing; it results in either deletion of exon 17 or deletion of exon 18 (ref. 8). The pedigrees of families 1 9 are shown in **Figure 1**.

[1]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, UK. [2]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratories, University of Cambridge, Cambridge CB1 8RN, UK. [3]Department of Medical Genetics, St. Mary's Hospital, Manchester M13 0JH, UK. [4]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton S016 6YA, UK. [5]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK. Correspondence should be addressed to N.R. (nazneen.rahman@icr.ac.uk).

Received 10 August; accepted 11 September; published online 8 October 2006; doi:10.1038/ng1902

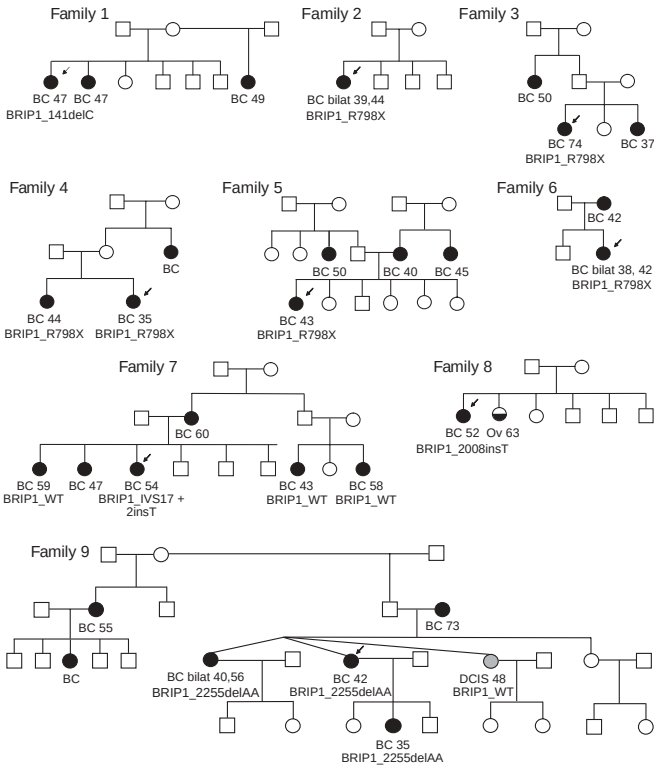


Figure 1 Abridged pedigrees of nine breast cancer families with *BRIP1* mutations. Individuals screened for *BRIP1* mutations are indicated by arrows. Individuals with breast cancer are shown as filled circles, with the age at diagnosis given underneath. An individual with ductal carcinoma *in situ* but no invasive cancer is shown as a shaded circle. If the individual had metachronous bilateral breast cancer, two ages are given. Other cancers or medical conditions are not shown. Samples were not available from individuals with breast cancer that are not genotyped. The *BRIP1* mutation in each family is given in **Table 1** and listed below the individual. BC, breast cancer; BC bilat, bilateral breast cancer; Ov, ovarian cancer; DCIS, ductal carcinoma *in situ*; BRIP1 WT, *BRIP1* mutation absent. We obtained informed consent from all families, and the research was approved by the London Multicentre Research Ethics Committee (MREC/01/2/18).

there was limited evidence of linkage of *BRIP1* truncating mutations with breast cancer in the *BRIP1* positive pedigrees (**Fig. 1**). This is the typical, and expected, pattern of low penetrance susceptibility alleles[6,7]. On the basis of the population frequency and breast cancer risk derived from our study, *BRIP1* mutations have an estimated breast cancer attributable fraction of 0.20% (95% c.i. = 0.04% 0.44%) in the UK.

It has previously been suggested that certain *BRIP1* missense variants may confer susceptibility to breast cancer[2,3]. We identified 24 nonsynonymous *BRIP1* missense variants, of which seven were present in both affected individuals and controls, eight were present exclusively in affected individuals and nine were present exclusively in controls (**Supplementary Table 2** online). The P919S variant had allele frequencies of 40.3% in affected individuals and 39.3% in controls ($P = 0.43$). The other 23 variants were each observed in $<1.5\%$ of the samples, with no significant difference in the frequency of any single variant or in their combined frequency between affected individuals and controls ($P = 0.29$). There was also no significant difference between affected individuals and controls in the *in silico* predicted effect on protein function or the position of missense variants within the gene (**Supplementary Methods**). These data indicate that the majority of *BRIP1* missense variants are not associated with a risk of breast cancer comparable to that conferred by truncating variants. However, we cannot exclude the possibility that a small number of specific missense alterations confer susceptibility to breast cancer. Notable in this regard is P47A, which was first reported in an individual with early onset breast cancer and a strong family history of breast and ovarian cancer[2]. This variant alters a highly conserved residue and has been shown to abolish BRIP1 helicase activity[2,3]. It was therefore considered likely that the presence of P47A was causally related to the cancer clustering in the family. However, we identified P47A in four affected individuals and four controls ($P = 0.48$), indicating that, despite the deleterious effect on BRIP1 function *in vitro*, it is unlikely to confer a risk of breast cancer similar to that of truncating mutations.

While we were conducting this study, biallelic inactivating *BRIP1* mutations were reported as the cause of Fanconi anemia complementation group J (FA J)[8–10]. Three of the six truncating *BRIP1* mutations we identified were also reported in Fanconi anemia patients. This includes the commonest *BRIP1* mutation in FA J cases, R798X, which we identified in five separate breast cancer families and one control. None of the FA J families were reported to have a strong family history of breast cancer consistent with the modest increased risk of breast cancer conferred by *BRIP1* mutations. Moreover, no FA J case with P47A has been reported, further suggesting that this variant may not be associated with the same cancer risks as truncating mutations. Of note, biallelic mutations of the breast cancer susceptibility gene *BRCA2* have been shown to cause Fanconi anemia complementation group D1 (FA D1)[11].

There are currently 11 known Fanconi anemia genes, and at least one additional gene (underlying complementation group I) awaits identification[12]. Epidemiological surveys of relatives of individuals with Fanconi anemia from all complementation groups combined have not provided evidence of an association with breast cancer[12,13]. However, FA D1 and FA J are rare subtypes, and therefore the risks of breast cancer they confer could easily be obscured in studies of all Fanconi subtypes together. Indeed, we have previously analyzed the genes underlying FA A, FA C, FA D2, FA E, FA F and FA G (which together account for over 90% of Fanconi anemia cases) in 88 familial breast cancer cases, and we did not identify any truncating mutations[14]. More extensive mutational surveys of FA genes in individuals with breast cancer are now indicated. Notably, however, 8 of the 11 known FA genes encode proteins that form a nuclear core complex that mediates the monoubiquitination of FANCD2. In contrast, *BRCA2* and *BRIP1* are Fanconi anemia genes encoding proteins that function downstream of FANCD2 (ref. 12).

Despite the functional and genetic similarities between *BRCA2* and *BRIP1*, there are some interesting differences in the phenotypes associated with mutations in these genes. Biallelic *BRCA2* mutations confer a high risk of childhood solid and hematological cancers[15], whereas, to date, only one cancer has been reported in an individual with FA J who has biallelic *BRIP1* mutations[8–10]. Monoallelic *BRCA2* mutations confer high risks of breast cancer, whereas monoallelic *BRIP1* mutations confer more modest risks, similar to truncating variants of *CHEK2* and *ATM*[6,7]. The biological explanations for the differences in cancer risk between *BRIP1* and *BRCA2* are currently unclear.

Five other genes implicated in DNA repair are known to confer susceptibility to breast cancer: *TP53*, *BRCA1*, *BRCA2*, *CHEK2* and *ATM*. These genes, together with *BRIP1*, still account

only for a minority of the familial aggregation of breast cancer. However, their close functional interactions suggest that other genes involved in DNA repair processes may also be involved in breast cancer susceptibility.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the participating families who were recruited to the study by the Familial Breast Cancer Collaboration (UK), which includes the following contributors: A. Ardern-Jones, J. Berg, A. Brady, N. Bradshaw, C. Brewer, G. Brice, B. Bullman, J. Campbell, B. Castle, R. Cetnarsryj, C. Chapman, C. Chu, N. Coates, T. Cole, R. Davidson, A. Donaldson, H. Dorkins, F. Douglas, D. Eccles, R. Eeles, F. Elmslie, D.G. Evans, S. Goff, S. Goodman, D. Goudie, J. Gray, L. Greenhalgh, H. Gregory, N. Haites, S.V. Hodgson, T. Homfray, R.S. Houlston, L. Izatt, L. Jeffers, V. Johnson-Roffey, F. Kavalier, C. Kirk, F. Lalloo, I. Locke, M. Longmuir, J. Mackay, A. Magee, S. Mansour, Z. Miedzybrodzka, J. Miller, P. Morrison, V. Murday, J. Paterson, M. Porteous, N. Rahman, M. Rogers, S. Rowe, S. Shanley, A. Saggar, G. Scott, L. Side, L. Snadden, M. Steel, M. Thomas and S. Thomas. We thank A. Hall and E. Mackie for coordination of sample collection. We are grateful for the support of the Daniel Faulkner Trust and the Geoffrey Berger Charitable Trust. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. P. Kelly is supported by a grant from the Breast Cancer Campaign. M.A., B.N. and A.E. are supported by US Army Medical Research and Material Command grant #W81XWH-05-1-0204. This research was supported by the Institute of Cancer Research and Cancer Research UK.

AUTHOR CONTRIBUTIONS

The study was designed by N.R. and M.R.S. The molecular analyses were performed by S.S., A.R., P.K., R.B., T.C., H.J., M.A. and K.S. under the direction of N.R. The statistical analyses were performed by D.T., A.E., B.N. and L.M. under the direction of D.F.E. The familial collections were initiated by G.E. and D.E. and were collected by the Breast Cancer Susceptibility Collaboration (UK). The manuscript was written by N.R. and M.R.S.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Antoniou, A.C. & Easton, D.F. *Genet. Epidemiol.* **21**, 1 18 (2001).
2. Cantor, S.B. *et al. Cell* **105**, 149 160 (2001).
3. Cantor, S. *et al. Proc. Natl. Acad. Sci. USA* **101**, 2357 2362 (2004).
4. Peng, M. *et al. Oncogene* **25**, 2245 2253 (2006).
5. Cantor, S.B. & Andreassen, P.R. *Cell Cycle* **5**, 164 167 (2006).
6. Meijers-Heijboer, H. *et al. Nat. Genet.* **31**, 55 59 (2002).
7. Renwick, A. *et al. Nat. Genet.* **38**, 873 875 (2006).
8. Levitus, M. *et al. Nat. Genet.* **37**, 934 935 (2005).
9. Levran, O. *et al. Nat. Genet.* **37**, 931 933 (2005).
10. Litman, R. *et al. Cancer Cell* **8**, 255 265 (2005).
11. Howlett, N.G. *et al. Science* **297**, 606 609 (2002).
12. Taniguchi, T. & D'Andrea, A.D. *Blood* **107**, 4223 4233 (2006).
13. Swift, M. *et al. J. Natl. Cancer Inst.* **65**, 863 867 (1980).
14. Seal, S. *et al. Cancer Res.* **63**, 8596 8599 (2003).
15. Reid, S. *et al. J. Med. Genet.* **42**, 147 151 (2005).

nature genetics

PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene

Nazneen Rahman[1], Sheila Seal[1], Deborah Thompson[2], Patrick Kelly[1], Anthony Renwick[1], Anna Elliott[1], Sarah Reid[1], Katarina Spanova[1], Rita Barfoot[1], Tasnim Chagtai[1], Hiran Jayatilake[1], Lesley McGuffog[2], Sandra Hanks[1], D Gareth Evans[3], Diana Eccles[4], The Breast Cancer Susceptibility Collaboration (UK), Douglas F Easton[2] & Michael R Stratton[1,5]

PALB2 interacts with BRCA2, and biallelic mutations in *PALB2* (also known as *FANCN*), similar to biallelic *BRCA2* mutations, cause Fanconi anemia. We identified monoallelic truncating *PALB2* mutations in 10/923 individuals with familial breast cancer compared with 0/1,084 controls ($P = 0.0004$) and show that such mutations confer a 2.3-fold higher risk of breast cancer (95% confidence interval (c.i.) = 1.4 3.9, $P = 0.0025$). The results show that *PALB2* is a breast cancer susceptibility gene and further demonstrate the close relationship of the Fanconi anemia DNA repair pathway and breast cancer predisposition.

PALB2 (for 'partner and localizer of BRCA2') encodes a recently discovered protein that interacts with BRCA2, is implicated in its nuclear localization and stability and is required for some functions of BRCA2 in homologous recombination and double strand break repair[1]. In a paper in this issue, we show that biallelic *PALB2* mutations are responsible for a subset of Fanconi anemia cases characterized by a phenotype similar to that caused by biallelic *BRCA2* mutations[2]. Prompted by these observations, we investigated whether monoallelic *PALB2* mutations confer susceptibility to breast cancer by sequencing the gene in individuals with breast cancer from familial breast cancer pedigrees that were negative for mutations in *BRCA1* and *BRCA2* and controls (**Supplementary Methods** online).

We identified truncating *PALB2* mutations in 10/923 (1.1%) independently ascertained individuals with familial breast cancer from separate families compared with 0/1,084 (0%) controls ($P = 0.0004$) (**Table 1** and **Fig. 1a**). Nine of the *PALB2* mutations were in the 908 families with female breast cancer only (1.0%). One occurred in the 15 families (6.7%) with cases of both female and male breast cancer ($P = 0.15$). Although this observation requires further investigation, it suggests that *PALB2* mutations may confer a higher relative risk of male breast cancer than female breast cancer, and *BRCA2* mutations are known to confer a high relative risk of male breast cancer[3]. One proband with a *PALB2* mutation developed melanoma at 47 years of age in addition to breast cancer at 56 years. Apart from this individual, there were no other malignancies other than breast cancer in individuals with *PALB2* mutations. Two of four first degree affected relatives of probands with *PALB2* mutations also carried a *PALB2* mutation. This pattern of incomplete segregation in affected relatives is typical of susceptibility alleles that confer modestly increased risks and is similar to that reported in breast cancer families carrying *CHEK2*, *ATM* or *BRIP1* mutations[4–6].

Segregation analysis incorporating the information from controls and the full pedigrees of the affected individuals estimated the relative risk of *PALB2* mutations to be 2.3 (c.i. = 1.4 3.9, $P = 0.0025$). The relative risk for women under 50 years was 3.0 (95% c.i. = 1.4 5.5), and for women over 50 years it was 1.9 (95% c.i. = 0.8 3.7, $P = 0.35$ for difference in relative risk between the age groups). The median age at diagnosis of individuals with *PALB2* mutations was 46 years (interquartile range (IQR) = 40 51) compared with a median age at diagnosis of 49 years (IQR = 42 55) in individuals with breast cancer without *PALB2* mutations ($P = 0.24$ for difference). These data suggest that the risks of breast cancer associated with *PALB2* mutations may be age dependent, but additional studies will be required to address this question. There was no difference in the extent

Table 1 Cancer history and *PALB2* mutations identified through analyses of individuals with familial breast cancer and controls

Family	Cancer history and age of proband	Number of relatives with breast cancer	*PALB2* mutation	PALB2 alteration
1	Breast cancer, 32 years	2	2386G→T	G796X
2	Breast cancer, 51 years	2 female, 1 male	2982insT	A995fs
3	Breast cancer, 43 years	3	3113G→A	W1038X
4	Breast cancer, 49 years	4	3113G→A	W1038X
5	Breast cancer, 28 years	2	3116delA	N1039fs
6	Breast cancer, 50 years	2	3116delA	N1039fs
7	Breast cancer, 55 years	3	3116delA	N1039fs
8	Breast cancer, 42 years	3	3549C→G	Y1183X
9	Breast cancer, 56 years Melanoma, 47 years	3	3549C→G	Y1183X
10	Breast cancer, 40 years	3	3549C→G	Y1183X

The mutations identified in families 5 10 have previously been reported as causative in individuals with Fanconi anemia subtype N (ref. 2; none of the FA-N families are part of this study). The probands with identical mutations were from separately ascertained families that are not known to be related and are from different parts of the UK. The pedigrees of families 1 10 are shown in **Figure 1**. We did not find any truncating mutations in sequencing the full *PALB2* coding sequence from 1,084 controls.

[1]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. [2]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratories, University of Cambridge, UK. [3]Department of Medical Genetics, St. Mary's Hospital, Manchester M13 0JH, UK. [4]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton SO16 6YA, UK. [5]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, UK. Correspondence should be addressed to N.R (nazneen.rahman@icr.ac.uk).

Received 24 October; accepted 8 December; published online 31 December 2006; doi:10.1038/ng1959

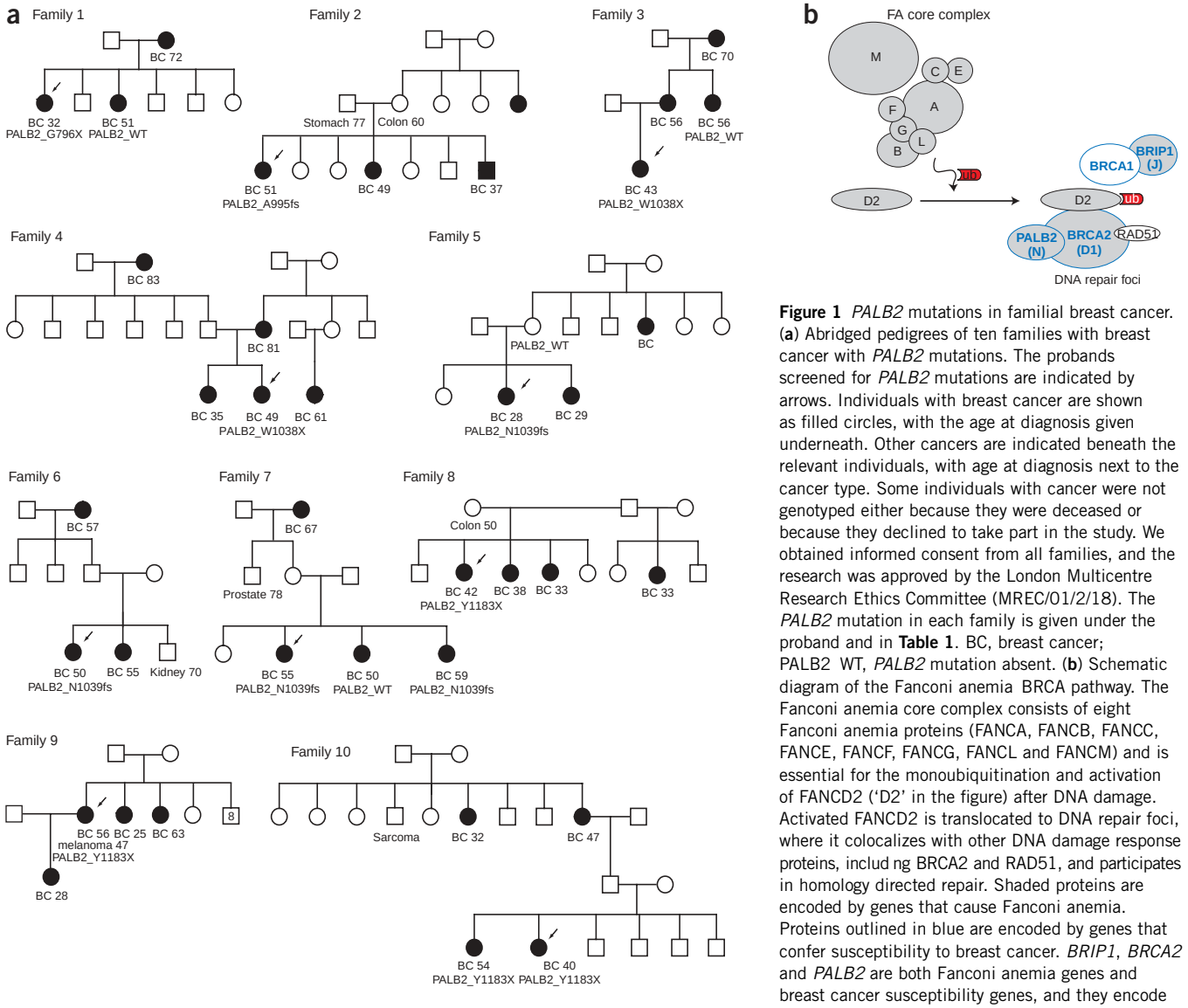


Figure 1 *PALB2* mutations in familial breast cancer. (**a**) Abridged pedigrees of ten families with breast cancer with *PALB2* mutations. The probands screened for *PALB2* mutations are indicated by arrows. Individuals with breast cancer are shown as filled circles, with the age at diagnosis given underneath. Other cancers are indicated beneath the relevant individuals, with age at diagnosis next to the cancer type. Some individuals with cancer were not genotyped either because they were deceased or because they declined to take part in the study. We obtained informed consent from all families, and the research was approved by the London Multicentre Research Ethics Committee (MREC/01/2/18). The *PALB2* mutation in each family is given under the proband and in **Table 1**. BC, breast cancer; PALB2 WT, *PALB2* mutation absent. (**b**) Schematic diagram of the Fanconi anemia BRCA pathway. The Fanconi anemia core complex consists of eight Fanconi anemia proteins (FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FANCM) and is essential for the monoubiquitination and activation of FANCD2 ('D2' in the figure) after DNA damage. Activated FANCD2 is translocated to DNA repair foci, where it colocalizes with other DNA damage response proteins, includ ng BRCA2 and RAD51, and participates in homology directed repair. Shaded proteins are encoded by genes that cause Fanconi anemia. Proteins outlined in blue are encoded by genes that confer susceptibility to breast cancer. *BRIP1*, *BRCA2* and *PALB2* are both Fanconi anemia genes and breast cancer susceptibility genes, and they encode proteins functioning downstream of FANCD2.

of familial clustering of breast cancer ($P = 0.69$) or in the probability of being a bilateral case ($P = 0.23$) in families with *PALB2* mutations compared with families without mutations. Assuming a conservative sensitivity of 90% for mutation detection, we estimate the breast cancer population attributable fraction of *PALB2* mutations to be 0.23% (95% c.i.: 0.072% 0.52%) and the percentage of the familial relative risk due to *PALB2* to be 0.24% (0.02% 1.16%).

We identified 50 nontruncating variants within the *PALB2* coding sequence, including 35 nonsynonymous and 15 synonymous variants (**Supplementary Table 1** online). There was no overall evidence that *PALB2* missense variants confer susceptibility to breast cancer, with 215 (23%) affected individuals and 265 (24%) controls carrying at least one nonsynonymous missense variant. Only four missense variants had an allele frequency greater than 1%, and there was no evidence that any of these were breast cancer susceptibility alleles. This result is consistent with the data from individuals with Fanconi anemia in which all reported *PALB2* mutations result in premature protein truncation[2,7].

Fanconi anemia is a genetically heterogeneous recessive condition that currently includes 13 subtypes, 12 of which have been attributed to distinct genes[2,8]. The known Fanconi anemia genes encode proteins that interact in an incompletely understood fashion to facilitate recognition and repair of DNA double strand breaks. A key process in the pathway involves eight of the known Fanconi anemia proteins forming a nuclear core complex that mediates monoubiquitination and activation of FANCD2. Activated FANCD2 is translocated to DNA repair foci, where it colocalizes with BRCA2 and other proteins that effect DNA repair by homologous recombination (**Fig. 1b**)[8].

Biallelic mutations of *BRCA2* and *PALB2* cause Fanconi anemia subtypes FA D1 and FA N, respectively[2,7,9]. The phenotypes associated with biallelic *BRCA2* and *PALB2* mutations are markedly similar to each other and differ from the other ten known Fanconi anemia genes. In particular, FA D1 and FA N are associated with high risks of solid childhood malignancies such as Wilms tumor and medulloblastoma, which occur very rarely in other subtypes[2,8,10]. Heterozygous muta tions in *BRIP1*, which encodes a BRCA1 interacting protein, also

confer an elevated risk of breast cancer[6], and biallelic *BRIP1* mutations cause Fanconi anemia subtype FA J[11,12]. However, FA J is associated with the classical Fanconi anemia phenotype, and there have not been any reports of individuals with FA J with a childhood solid tumor[11,12].

It is plausible that heterozygosity for mutations in other Fanconi anemia genes may also be involved in breast cancer susceptibility. However, epidemiological studies of relatives of individuals with Fanconi anemia have not demonstrated this, suggesting that breast cancer susceptibility is associated with only a subset of Fanconi anemia genes. This is consistent with the negative results of mutational screens of other Fanconi anemia genes in familial breast cancer cases[13]. The biological features that determine whether a Fanconi anemia gene is also a breast cancer predisposition gene are unknown. However, it is notable that the three Fanconi anemia genes currently associated with breast cancer susceptibility (*BRCA2*, *PALB2* and *BRIP1*) are not part of the Fanconi anemia core complex and are the only known Fanconi anemia genes that act downstream of FANCD2 (**Fig. 1b**).

We estimate that *PALB2* mutations are associated with an approxi mately twofold higher risk of female breast cancer. Therefore, despite the fact that PALB2 is functionally associated with BRCA2 and that biallelic mutations in both genes cause similar phenotypes, the increase in breast cancer risk associated with *PALB2* monoallelic mutations is clearly more modest than that conferred by *BRCA2* monoallelic mutations, which result in approximately a tenfold increase in risk. These differences in risk are reminiscent of those previously reported between *BRCA1* mutations, which also confer a greater than tenfold increase in risk of breast cancer, and mutations in *BRIP1*, which confer only a twofold increase in risk[6]. The explana tions for the apparent differences in risk associated with mutations in these genes, despite the close functional interactions between the proteins they encode, are currently unknown. Thus, our data provide further evidence of the close link between breast cancer susceptibility and the Fanconi anemia DNA repair pathway, but they also demon strate that the relationship is complex at both the phenotypic and molecular levels.

With the identification of *PALB2* as a new breast cancer predis position gene, a clearer picture of the genetic architecture of breast cancer susceptibility is emerging. *BRCA1* and *BRCA2* are likely to be the only major high penetrance breast cancer susceptibility genes (leading to more than a tenfold higher risk). Mutations in *TP53* also confer high risks of breast cancer but are much rarer[14]. These genes are characterized by multiple, rare, inactivating mutations that together account for approximately 15% 20% of the familial risk of the disease[14]. A similar mutation spectrum has now been identified in four additional genes that encode proteins that interact biologically with BRCA1, BRCA2 and/or p53. Three of these proteins, CHK2, ATM and BRIP1, interact with BRCA1, p53 or both (refs. 8,15). PALB2 is the first that interacts with BRCA2. However, compared with risks associated with mutations in *BRCA1*, *BRCA2* and *TP53*, the risks associated with mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* are much lower[4–6]. Moreover, inactivating mutations in each of these genes are rare, with fewer than 1% of the population being hetero zygotes. As such, the contribution of each gene to the familial risk of breast cancer is small. Collectively, however, they already account for ~2.3% of the overall familial relative risk. Thus, this class of susceptibility gene may make an appreciable contribution to breast cancer predisposition.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the participating families who were recruited to the study by the Familial Breast Cancer Collaboration (UK) which includes the following contributors: A. Ardern-Jones, J. Berg, A. Brady, N. Bradshaw, C. Brewer, G. Brice, B. Bullman, J. Campbell, B. Castle, R. Cetnarsryj, C. Chapman, C. Chu, N. Coates, T. Cole, R. Davidson, A. Donaldson, H. Dorkins, F. Douglas, D. Eccles, R. Eeles, F. Elmslie, D.G. Evans, S. Goff, S. Goodman, D. Goudie, J. Gray, L. Greenhalgh, H. Gregory, N. Haites, S.V. Hodgson, T. Homfray, R.S. Houlston, L. Izatt, L. Jeffers, V. Johnson-Roffey, F. Kavalier, C. Kirk, F. Lalloo, I. Locke, M. Longmuir, J. Mackay, A. Magee, S. Mansour, Z. Miedzybrodzka, J. Miller, P. Morrison, V. Murday, J. Paterson, M. Porteous, N. Rahman, M. Rogers, S. Rowe, S. Shanley, A. Saggar, G. Scott, L. Side, L. Snadden, M. Steel, M. Thomas and S. Thomas. We thank A. Hall and E. Mackie for coordination of sample collection. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. P.K. is supported by a grant from the Breast Cancer Campaign. A.E. is supported by US Army Medical Research and Material Command grant W81XWH-05-1-0204. This research was supported by donations from the Geoffrey Berger and Daniel Falkner Charitable Trusts, by the Institute of Cancer Research and by Cancer Research UK.

AUTHOR CONTRIBUTIONS

The study was designed by N.R. and M.R.S. The molecular analyses were performed by S.S., P.K., A.R., S.R., K.S., R.B., T.C., H.J. and S.H. under the direction of N.R. The statistical analyses were performed by D.T., A.E. and L.M. under the direction of D.F.E. The familial collections were initiated by D.G.E. and D.E. and were collected by the Breast Cancer Susceptibility Collaboration (UK). The manuscript was written by N.R. and M.R.S.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Xia, B. *et al. Mol. Cell* **22**, 719 729 (2006).
2. Reid, S. *et al. Nat. Genet.* advance online publication 31 December 2006 (doi:10.1038/ng1947).
3. The Breast Cancer Linkage Consortium. *J. Natl. Cancer Inst.* **91**, 1310 1316 (1999).
4. Meijers-Heijboer, H. *et al. Nat. Genet.* **31**, 55 59 (2002).
5. Renwick, A. *et al. Nat. Genet.* **38**, 873 875 (2006).
6. Seal, S. *et al. Nat. Genet.* **38**, 1239 1241 (2006).
7. Xia, B. *et al. Nat. Genet.* advance online publication 31 December 2006 (doi: 10.1038/ng1942).
8. Taniguchi, T. & D'Andrea, A.D. *Blood* **107**, 4223 4233 (2006).
9. Howlett, N.G. *et al. Science* **297**, 606 609 (2002).
10. Reid, S. *et al. J. Med. Genet.* **42**, 147 151 (2005).
11. Levitus, M. *et al. Nat. Genet.* **37**, 934 935 (2005).
12. Levran, O. *et al. Nat. Genet.* **37**, 931 933 (2005).
13. Seal, S. *et al. Cancer Res.* **63**, 8596 8599 (2003).
14. Antoniou, A.C. & Easton, D.F. *Oncogene* **25**, 5898 5905 (2006).
15. Shiloh, Y. *Trends Biochem. Sci.* **31**, 402 410 (2006).

nature genetics

Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants

Wellcome Trust Case Control Consortium[1] & The Australo-Anglo-American Spondylitis Consortium[1]

We have genotyped 14,436 nonsynonymous SNPs (nsSNPs) and 897 major histocompatibility complex (MHC) tag SNPs from 1,000 independent cases of ankylosing spondylitis (AS), autoimmune thyroid disease (AITD), multiple sclerosis (MS) and breast cancer (BC). Comparing these data against a common control dataset derived from 1,500 randomly selected healthy British individuals, we report initial association and independent replication in a North American sample of two new loci related to ankylosing spondylitis, *ARTS1* and *IL23R*, and confirmation of the previously reported association of AITD with *TSHR* and *FCRL3*. These findings, enabled in part by increased statistical power resulting from the expansion of the control reference group to include individuals from the other disease groups, highlight notable new possibilities for autoimmune regulation and suggest that *IL23R* may be a common susceptibility factor for the major 'seronegative' diseases.

Genome wide association scans are currently revealing a number of new genetic variants for common diseases[1–11]. We have recently completed the largest and most comprehensive scan conducted to date, involving genome wide association studies of 2,000 individuals from each of seven common disease cohorts and 3,000 common control individuals using a dense panel of >500,000 markers[12]. In parallel with this scan, we conducted a study of 5,500 independent individuals with a genome wide set of nonsynonymous coding variants, an approach that has recently yielded new findings about type 1 diabetes and Crohn's disease and that has been proposed as an efficient complementary approach to whole genome scans[13–15]. Here we report several new replicated associations in our scan of nsSNPs in 1,500 shared controls and 1,000 individuals from each of four different diseases: ankylosing spondylitis, AITD (of which all had Graves' disease), breast cancer and multiple sclerosis.

RESULTS

Initial genotyping was carried out with a custom made Infinium array (Illumina) and involved 14,436 nsSNPs (assays were synthesized for 16,078 nsSNPs). At the inception of the study, this comprised the complete set of experimentally validated nsSNPs with minor allele frequency (MAF) > 1% in western European samples. In addition, because three of the diseases were of autoimmune etiology, we also typed a dense set of 897 SNPs throughout the MHC that, together with 348 nsSNPs in this region, provided comprehensive tag SNP coverage ($r^2 \geq 0.8$ with all SNPs in ref. 16). Finally, 103 SNPs were typed in pigmentation genes specifically designed to differentiate between population groups. Similar to those from previous studies, our data revealed that detailed assessment of initial data is critical to the process of association inference, as biases in genotype calling lead to inflation of false positive rates[12,17]. This inflation is exaggerated in nsSNP data, because nsSNPs tend to have lower allele frequencies than otherwise anonymous genomic SNPs, and genotype calling is often most difficult for rare alleles. If only cursory filtering had been applied in the present case, numerous false positives would have emerged (**Supplementary Figs. 1 4** online). **Table 1** shows the total number of SNPs and individuals remaining after genotype and sample quality control procedures (see Methods).

Association with the MHC

The strongest associations observed in the study were between SNPs in the MHC region and the three autoimmune diseases studied ankylosing spondylitis, AITD and MS with P values of $<10^{-20}$ for each disease (**Fig. 1**). No association of the MHC was seen with breast cancer ($P > 10^{-4}$ across the region). For each of the autoimmune diseases, the maximum signal was centered around the known HLA associated genes (for example, those encoding HLA B in ankylosing spondylitis, HLA DRB1 in MS and the MHC class I and class II molecules in AITD), but in all cases, it extended far beyond the specific associated haplotype(s). For example, in ankylosing spondylitis, association was observed at $P < 10^{-20}$ across ~1.5 Mb. Given the well known strong effect of HLA B27 variant on the probability of developing ankylosing spondylitis (odds ratio 100 200 in most popu lations), the extent of this association signal reflects that with such large effects, even very distant SNPs in modest linkage disequilibrium (LD) will show indirect evidence for association. Strong signals like these may also cloud the evidence for additional HLA loci[18]. Disen tangling similar patterns of association within the MHC has proven extremely challenging in the past and will be addressed in future studies of these data. Here we focus specifically on the nsSNP results.

[1]The complete lists of participants and affiliations appear at the end of the article. Correspondence should be addressed to L.R.C. (lcardon@fhcrc.org) or D.M.E. (davide@well.ox.ac.uk).

Received 17 July; accepted 17 September; published online 21 October 2007; doi:10.1038/ng.2007.17

Table 1 Number of individuals and SNPs tested in each cohort

	Cohort				
	AS	AITD	BC	MS	58C
Males	610	138	0	271	732
Females	312	762	1,004	704	734
Number of SNPs genotyped	15,436	15,436	15,436	15,436	15,436
SNPs with low GC score	783	816	771	802	796
SNPs with low genotyping	133	206	124	218	186
Monomorphic SNPs	1,842	1,829	1,854	1,810	1,687
SNPs with HW $P < 10^{7}$[a]	129	74	104	97	132
Differences in missing rate $P < 10^{4}$	51	101	172	309	n/a
'Manual' exclusions	33	33	33	33	33
Total number of SNPs tested	12,701	12,572	12,577	12,374	

[a]Only SNPs with HW $P < 10^{7}$ in the 1958 birth cohort (58C) control group were excluded from analyses.

Association with nsSNPs

A major advantage of the Wellcome Trust Case Control Consortium (WTCCC) design is the availability of multiple disease cohorts that are similar in terms of ancestry and that have been typed on the same genetic markers[12,17]. Assuming that each disease has at least some unique genetic loci, we hypothesized that combining the other three case groups with the controls for the 1958 birth cohort (58C)[19] would increase power to detect association. For each disease, we therefore conducted two primary analyses: first, we tested nsSNP associations for each disease against the controls in the 58C; and second, we tested the same associations for each disease against an expanded reference group comprising the combined cases from the other three disease groups plus individuals from the 58C. A similar set of analyses was conducted for each of the autoimmune disorders against a reference group comprising 58C controls and individuals with breast cancer, but the results were very similar to those for the fully expanded groups, so here we describe the larger sample (**Supplementary Table 1** online). In addition, because it is possible that different autoimmune diseases share similar genetic etiologies, we also compared a combined ankylosing spondylitis, AITD and MS group (immune cases) against the combined set of individuals with breast cancer and 58C controls. All of our analyses are reported without regard to specific treatment of population structure, as the degree of structure in our final genotype data is not severe (Genomic Control[20] $\lambda = 1.07$ 1.13 in the 58C only datasets; $\lambda = 1.03$ 1.06 in the expanded reference group comparisons; **Table 2**), consistent with our recent findings from 17,000 UK individuals involving the same controls[12].

nsSNP association results (excluding the MHC region) for each of the four disease groups against the 58C controls are shown in **Figure 2** and **Table 3**. Two SNPs on chromosome 5 reached a high level of statistical significance for ankylosing spondylitis (rs27044: $P = 1.0 \times 10^{-6}$; rs30187: $P = 3.0 \times 10^{-6}$). This level of significance exceeds the 10^{-5} 10^{-6} thresholds advocated for gene based scans[21], as well as the oft used Bonferroni correction at $P < 0.05$ (see refs. 12,21 for a discussion of genome wide association significance). Both of these markers reside in the gene *ARTS1* (*ERAAP*, *ERAP1*), which encodes a type II integral transmembrane aminopeptidase with diverse immunological functions. Four additional SNPs show significance at $P < 10^{-4}$, with an increasing number of possible associations at more modest significance levels. Several of the more strongly associated SNPs, and others in the same genes, have been previously associated with these particular diseases, and for yet others there exists functional evidence of involvement in these particular conditions. Among these are SNPs in *FCRL3* and *FCRL5* in the case of AITD, *IL23R* in the case of ankylosing spondylitis, *MEL18* in the case of breast cancer and *IL7R* for MS. The complete list of single marker association results is provided in **Supplementary Table 1**.

The results of analyses involving the expanded reference group are presented in **Supplementary Figure 5** online and **Supplementary**

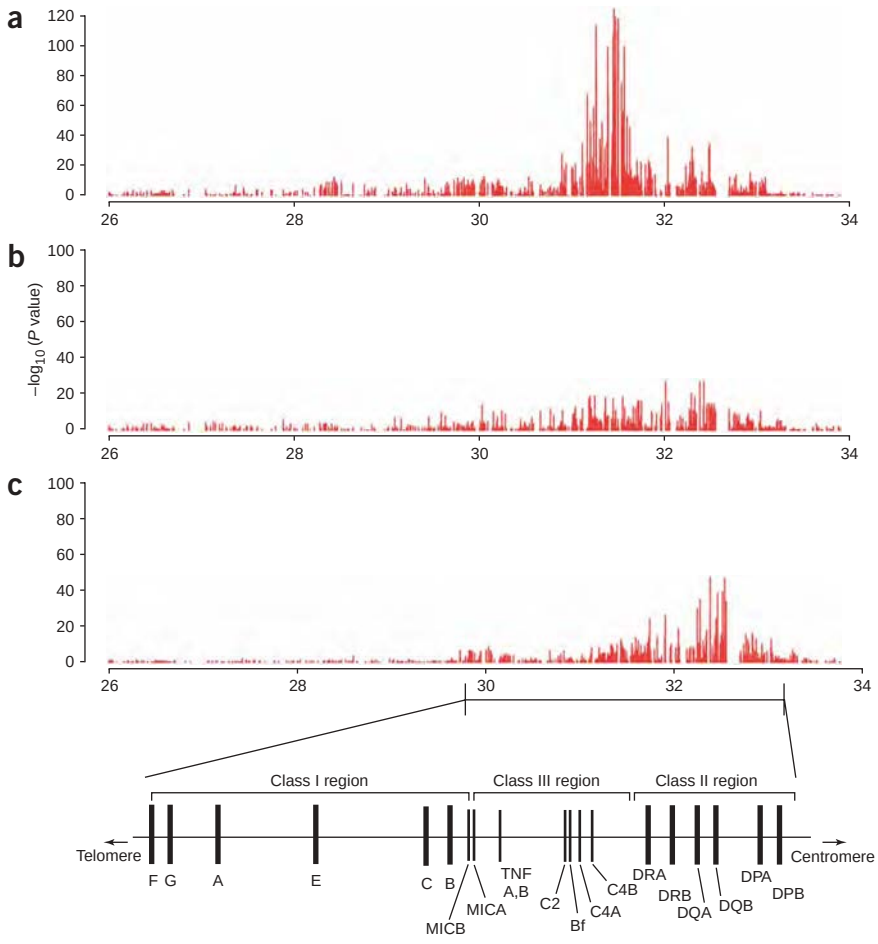


Figure 1 Minus $\log_{10}$ *P* values for the Armitage test of trend for MHC association with ankylosing spondylitis (**a**), auto mmune thyroid disease (**b**) and multiple sclerosis (**c**). Note in particular how evidence for association extends along very long regions of the MHC, reflecting statistical power to detect association even when linkage disequilibrium amongst SNPs is relatively low or when there exists the possibility of multiple disease predisposing loci.

Table 2 Estimates of λ for single and combined cohorts

		λ
Single cohort	AS cases versus 58C	1.07
	AITD cases versus 58C	1.12
	BC cases versus 58C	1.13
	MS cases versus 58C	1.12
Mixed cohorts	AS cases versus all others	1.03
	AITD cases versus all others	1.05
	BC cases versus all others	1.04
	MS cases versus all others	1.06
	IMMUNE cases versus BC and 58C	1.04

Table 1. Many of the SNPs that showed moderate to strong evidence for association in the initial analysis had substantially greater signifi cance when the larger reference group was used. Notably, these included the SNPs rs27044 ($P = 4.0 \times 10^{-8}$) and rs30187 ($P = 2.1 \times 10^{-7}$) in *ARTS1*, as well as several other variants in this gene. A second SNP, rs7302230 in the gene encoding calsyntenin 3 on chromosome 12, showed substantially stronger evidence for associa tion in the expanded reference group analysis ($P = 5.3 \times 10^{-7}$) relative to the 58C only results ($P = 1.1 \times 10^{-4}$). Results of the expanded group also showed elevated results for several SNPs that did not appear exceptional in the original (non combined) analyses, including SNPs in several candidate genes such as those encoding sialoadhesin[22] and complement receptor 1 for ankylosing spondylitis, *PIK3R2* for MS, and *C8B*, *IL17R* and *TYK2* in the combined autoimmune disease analysis. SNP rs3783941 in the gene *TSHR*, encoding the thyroid stimulating hormone receptor, emerged as among the most significant in the expanded reference group analyses of AITD ($P = 2.1 \times 10^{-5}$). Several polymorphisms in *TSHR* have previously been associated with Graves' disease[23,24]. This known association did not reach even the modest significance level of 10^{-3} in the original analyses, but the addition of 3,000 further reference samples delineated it from the background noise and further sup ports the original independent report.

ARTS1 association confirmed in an independent cohort

To validate the most exceptional findings from the initial study, we genotyped the *ARTS1*, *CLSTN3* and *LNPEP* SNPs in 471 independent ankylosing spondylitis cases (**Table 4**) and 625 new controls (all self identified North American Caucasian). The data strongly suggest that the *ARTS1* associa tion is genuine. All *ARTS1* nsSNPs revealed independent replication in the same direction of effect, with replication significance levels ranging from 4.7×10^{-4} to 5.1×10^{-5}. When combined with the original samples, the results showed strong evidence for association with ankylosing spondylitis ($P = 1.2 \times 10^{-8}$ to 3.4×10^{-10}). The population attributable risk[25] contributed by the most strongly asso ciated marker in the North American dataset (rs2287987) was 26%.

Association was also confirmed with marker rs2303138 in the *LNPEP* gene, which lies 127 kb 3′ of *ARTS1*. This marker was in strong LD with *ARTS1* markers ($D' = 1$, rs27044 rs2303138). We tested the interdependence of the *ARTS1* and *LNPEP* associations using conditional logistic regression. The remaining association at *LNPEP* was weak after controlling for *ARTS1* ($P = 0.01$), whereas the association at *ARTS1* remained strong after controlling for *LNPEP* ($P = 2.7 \times 10^{-6}$), suggesting that the *LNPEP* association may only be secondary to LD, with a true association at *ARTS1*.

No association was seen with *CLSTN3* in the confirmation set. The US controls showed the same allele frequency as the UK controls (5%), but the allele frequency in the US cases was less than that of the UK cases (6% versus 8%), suggesting no association in the US samples and substantially reducing the significance of the combined data. Calystenin 3 is a postsynaptic neuronal membrane protein and is an unlikely candidate for involvement in inflammatory arthritis. The failure to replicate this association suggests that our replication sample size was insufficient to detect the modest effect or that it was a false positive in the initial scan.

IL23R variants confer risk of ankylosing spondylitis

The *IL23R* variant rs11209026, although not notable in the initial nsSNP scan ($P = 1.7 \times 10^{-3}$), was of particular interest, as it has recently been associated with both Crohn's disease[26,27] and psoriasis[28], conditions that commonly co occur with ankylosing spondylitis. To better define this association, seven additional SNPs in *IL23R* were genotyped in the same 1,000 British ankylosing spondylitis cases and 1,500 58C controls as well as the North American Caucasian replica tion samples (**Table 4**). In the WTCCC dataset, we observed strong association in seven of eight genotyped SNPs ($P \leq 0.008$, including the original nsSNP rs11209026), with the strongest association at rs11209032 ($P = 2.0 \times 10^{-6}$). In the replication dataset, we noted association with all genotyped SNPs ($P \leq 0.04$), with peak association with marker rs10489629 ($P = 4.2 \times 10^{-5}$). In the combined dataset,

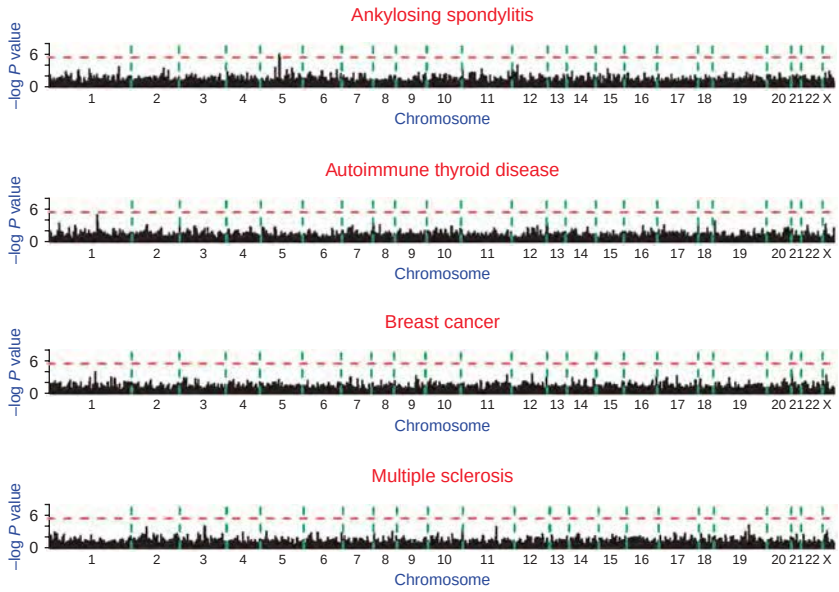


Figure 2 Minus $\log_{10}$ *P* values for the Armitage test of trend for genome wide association scans of ankylosing spondylitis, autoimmune thyroid disease, breast cancer and multiple sclerosis. The spacing between SNPs on the plot is uniform and does not reflect distances between the SNPs. The vertical dashed lines reflect chromosomal boundaries. The horizontal dashed lines display the cutoff of $P = 10^{-6}$. Note that SNPs within the MHC are not included in this diagram.

the strongest association observed was with SNP rs11209032 (odds ratio 1.3, 95% confidence interval 1.2 1.4, $P = 7.5 \times 10^{-9}$). The attributable risk for this marker in the replication cohort is 9%. Conditional logistic regression analyses did not indicate a single primary disease associated marker; residual association remained after we controlled for association at the remaining SNPs. Considering only individuals with ankylosing spondylitis who self reported as not having inflammatory bowel disease ($n = 1{,}066$) the association remained strong and was still strongest at rs11209032 ($P = 6.9 \times 10^{-7}$), indicating that there is a primary association with ankylosing spondylitis and that the observed association was not due to coexistent clinical inflammatory bowel disease.

In contrast to the pleiotropic effects of *IL23R*, the *ARTS1* association evidence seems confined to ankylosing spondylitis. We genotyped

Table 3 nsSNPs outside the MHC that meet a point-wise significance level of $P < 10^{-3}$ for the Cochran-Armitage test for trend

Disease	SNP	Chromosome	Position (bp)	MAF	OR	χ^2	*P* value	Gene
AS	rs696698	1	74777462	0.04	1.84	11.13	8.5×10^{-4}	*C1orf173*
	rs10494217	1	119181230	0.17	0.77	11.62	6.5×10^{-4}	*TBX15*
	rs2294851	1	206966279	0.13	0.73	13.55	2.3×10^{-4}	*HHAT*
	rs8192556	2	182368504	0.01	0.45	12.24	4.7×10^{-4}	*NEUROD1*
	rs16876657	5	78645930	0.02	3.10	13.05	3.0×10^{-4}	*JMY*
	rs27044	5	96144608	0.34	1.40	23.90	1.0×10^{-6}	*ARTS 1*
	rs17482078	5	96144622	0.17	0.76	13.55	2.3×10^{-4}	*ARTS 1*
	rs10050860	5	96147966	0.18	0.75	14.87	1.1×10^{-4}	*ARTS 1*
	rs30187	5	96150086	0.40	1.33	21.82	3.0×10^{-6}	*ARTS 1*
	rs2287987	5	96155291	0.18	0.75	14.31	1.6×10^{-4}	*ARTS 1*
	rs2303138	5	96376466	0.10	1.58	19.41	1.1×10^{-5}	*LNPEP*
	rs11750814	5	137528564	0.16	0.77	10.99	9.1×10^{-4}	*BRD8*
	rs11959820	5	149192703	0.02	0.49	12.41	4.3×10^{-4}	*PPARGC1B*
	rs907609	11	1813846	0.13	0.76	10.91	9.5×10^{-4}	*SYT8*
	rs3740691	11	47144987	0.29	0.80	11.86	5.7×10^{-4}	*ZNF289*
	rs11062385	12	297836	0.24	0.79	11.82	5.9×10^{-4}	*JARID1A*
	rs7302230	12	7179699	0.08	1.57	14.97	1.1×10^{-4}	*CLSTN3*
AITD	rs10916769	1	20408244	0.17	0.76	12.10	5.0×10^{-4}	*FLJ32784*
	rs6427384	1	154321955	0.18	1.43	18.97	1.3×10^{-5}	*FCRL5*
	rs2012199	1	154322098	0.17	1.35	13.18	2.8×10^{-4}	*FCRL5*
	rs6679793	1	154327170	0.22	1.33	14.69	1.3×10^{-4}	*FCRL5*
	rs7522061	1	154481463	0.47	1.25	13.78	2.1×10^{-4}	*FCRL3*
	rs1047911	2	74611433	0.15	1.34	11.24	8.0×10^{-4}	*MRPL53*
	rs7578199	2	241912838	0.26	1.26	11.53	6.9×10^{-4}	*HDLBP*
	rs3748140	8	9036429	0.00	0.28	11.44	7.2×10^{-4}	*PPP1R3B*
	rs1048101	8	26683945	0.42	0.82	10.98	9.2×10^{-4}	*ADRA1A*
	rs7975069	12	132389146	0.30	0.80	12.06	5.2×10^{-4}	*ZNF268*
	rs2271233	17	6644845	0.07	0.94	11.32	7.7×10^{-4}	*TEKT1*
	rs2856966	18	897710	0.19	0.76	14.00	1.8×10^{-4}	*ADCYAP1*
	rs7250822	19	2206311	0.04	1.97	13.83	2.0×10^{-4}	*AMH*
	rs2230018	23	44685331	0.14	1.41	11.55	6.8×10^{-4}	*UTX*
BC	rs4255378	1	151919300	0.48	1.25	14.70	1.3×10^{-4}	*MUC1*
	rs2107732	7	44851218	0.10	1.40	10.96	9.3×10^{-4}	*CCM2*
	rs4986790	9	117554856	0.07	1.54	11.46	7.1×10^{-4}	*TLR4*
	rs2285374	11	118457383	0.38	0.82	12.25	4.7×10^{-4}	*VPS11*
	rs7313899	12	54231386	0.03	2.10	13.02	3.1×10^{-4}	*OR6C4*
	rs2879097	17	34143085	0.20	0.78	11.73	6.1×10^{-4}	*MEL18*
	rs2822558	21	14593715	0.13	0.73	13.87	2.0×10^{-4}	*ABCC13*
	rs2230018	23	44685331	0.14	1.40	12.14	4.9×10^{-4}	*UTX*
MS	rs17009792	2	74400978	0.02	0.44	14.41	1.5×10^{-4}	*SLC4A5*
	rs1132200	3	120633526	0.15	0.73	15.22	9.6×10^{-5}	*FLJ10902*
	rs6897932	5	35910332	0.23	0.80	11.04	8.9×10^{-4}	*IL7R*
	rs6470147	8	124517985	0.36	1.23	10.92	9.5×10^{-4}	*FLJ10204*
	rs3818511	10	134309378	0.24	1.28	12.84	3.4×10^{-4}	*INPP5A*
	rs11574422	11	67970565	0.02	2.82	14.64	1.3×10^{-4}	*LRP5*
	rs388706	19	49110533	0.48	1.22	11.19	8.2×10^{-4}	*ZNF45*
	rs1800437	19	50873232	0.17	0.74	16.11	6.0×10^{-5}	*GIPR*
	rs2281868	23	69451484	0.50	1.26	11.38	7.4×10^{-4}	*SAP102*

Table 4 Ankylosing spondylitis replication results

		UK cases				US cases				All cases			
Gene	SNP	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value
ARTS1	rs27044	0.34	0.27	1.40	1.0×10^{-6}								
ARTS1	rs17482078	0.17	0.22	0.76	2.3×10^{-4}	0.15	0.21	0.65	5.1×10^{-5}	0.16	0.22	0.70	1.2×10^{-8}
ARTS1	rs10050860	0.18	0.23	0.75	1.2×10^{-4}	0.15	0.22	0.66	8.8×10^{-5}	0.17	0.22	0.71	7.6×10^{-9}
ARTS1	rs30187	0.40	0.33	1.33	3.0×10^{-6}	0.41	0.35	1.30	0.00047	0.41	0.34	1.40	3.4×10^{-10}
ARTS1	rs2287987	0.18	0.22	0.75	1.6×10^{-4}	0.15	0.21	0.66	8.4×10^{-5}	0.17	0.22	0.71	1.0×10^{-8}
LNPEP	rs2303138	0.10	0.07	1.58	1.1×10^{-5}	0.11	0.09	1.40	0.018	0.11	0.07	1.48	1.1×10^{-6}
CLSTN3	rs7302230	0.08	0.05	1.57	1.1×10^{-4}	0.06	0.05	1.10	0.56	0.07	0.05	1.30	0.0039
IL23R	rs11209026	0.04	0.06	0.63	0.0017	0.038	0.06	0.63	0.014	0.04	0.06	0.63	4.0×10^{-6}
IL23R	rs1004819	0.35	0.30	1.20	0.0013	0.35	0.30	1.30	0.0045	0.35	0.30	1.20	1.1×10^{-5}
IL23R	rs10489629	0.43	0.45	0.90	0.062	0.39	0.47	0.72	4.2×10^{-5}	0.41	0.46	0.83	0.00011
IL23R	rs11465804	0.04	0.06	0.67	0.0019	0.049	0.06	0.68	0.04	0.04	0.06	0.68	0.0002
IL23R	rs1343151	0.30	0.34	0.85	0.0077	0.29	0.36	0.71	6.7×10^{-5}	0.30	0.34	0.80	1.0×10^{-5}
IL23R	rs10889677	0.36	0.31	1.20	0.00066	0.37	0.29	1.40	4.7×10^{-5}	0.36	0.31	1.30	1.3×10^{-6}
IL23R	rs11209032	0.38	0.32	1.30	2.0×10^{-6}	0.38	0.32	1.30	0.0013	0.38	0.32	1.30	7.5×10^{-9}
IL23R	rs1495965	0.49	0.44	1.20	0.0021	0.50	0.43	1.40	0.00019	0.49	0.44	1.20	3.1×10^{-6}

the five ankylosing spondylitis associated SNPs in 755 British Crohn's disease and 1,011 ulcerative colitis cases and 633 healthy controls. No association was seen with either ulcerative colitis or Crohn's disease (Armitage trend $P > 0.4$ for all markers).

FCRL3 confirmed in AITD pathogenesis

In addition to the ankylosing spondylitis replications, we attempted to confirm and extend the *FCRL3* association in AITD. The SNP rs7522061 in the *FCRL3* gene was recently reported to be associated with AITD[29] and two other autoimmune diseases, rheumatoid arthritis and systemic lupus erythematosus[30]. Our initial association evidence ($P = 2.1 \times 10^{-4}$) likely reflects the signal of the originally detected polymorphism, because the level of LD is high across this gene. In fact, the entire 1q21 q23 region (which includes another gene, *FCRL5*, flagged in our scan) has also been implicated in several autoimmune diseases, including psoriasis and multiple sclerosis[31,32].

On the basis of the original findings on 1q21 q23, the original cohort was increased from 1,000 to 2,500 Graves disease cases, and we used 2,500 controls from the 58C control set. We selected eight SNPs that tagged the *FCRL3* and *FCRL5* gene regions and typed them in all 5,000 samples using an alternative genotyping platform. SNP rs3761959, which tags rs7522061 and rs7528684 (previously associated with rheumatoid arthritis and Graves' disease), was associated with Graves' disease in this extended cohort (**Table 5**), confirming the original result. In total, three of the seven *FCRL3* SNPs showed some evidence for association ($P < 0.05$), with SNP rs11264798 showing the strongest association of the tag SNPs ($P = 4.0 \times 10^{-3}$). SNP rs6667109 in *FCRL5*, which tagged SNPs rs6427384, rs2012199 and rs6679793, all found to be weakly associated in the original study, showed little evidence of association in this extended cohort.

DISCUSSION

Our scan of nsSNPs has identified and validated two new genes (*ARTS1* and *IL23R*) associated with ankylosing spondylitis, confirmed and extended markers in the *TSHR* and *FCRL3* genes that have previously been associated with AITD, and provided a dense set of association data for AITD, ankylosing spondylitis and MS across the MHC region. The challenge now is to design functional studies that

Table 5 Autoimmune thyroid disease replication results

		Replication cohort				Combined cohort			
Gene	SNP	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value
FCRL3	rs3761959[a]	0.48	0.45	0.87	0.013	0.49	0.45	0.87	9.4×10^{-3}
FCRL3	rs11264794	0.42	0.45	1.10	0.079	0.42	0.46	1.12	0.013
FCRL3	rs11264793	0.27	0.24	0.87	0.029	0.26	0.24	0.90	0.044
FCRL3	rs11264798	0.44	0.49	1.18	4.0×10^{-3}	0.44	0.49	1.22	1.6×10^{-5}
FCRL3	rs10489678	0.19	0.20	1.04	0.58	0.20	0.20	1.04	0.43
FCRL3	rs6691569	0.28	0.29	1.02	0.75	0.29	0.29	1.00	0.93
FCRL3	rs2282284	0.062	0.058	0.92	0.015	0.062	0.058	0.93	0.47
FCRL5	rs6667109	0.17	0.16	0.93	0.38	0.18	0.15	0.85	7.7×10^{-2}

[a]This SNP tags the SNP rs7522061, which was flagged as associated with AITD in the WTCCC screen (P 2.1×10^{-4}).

will reveal how variation in these genes translates into physiological processes that influence disease risk.

From a functional perspective, *ARTS1* and *IL23R* represent excellent biological candidates for association with ankylosing spondylitis. The protein ARTS1 has two known functions, either of which may explain the association. First, within the endoplasmic reticulum, ARTS1 is involved in trimming peptides to the optimal length for MHC class I presentation[33,34]. Ankylosing spondylitis is primarily an HLA class I mediated autoimmune disease[35], with >90% of cases carrying the HLA B27 allele. How HLA B27 increases risk of developing ankylosing spondylitis is unknown, but if the association of *ARTS1* with the disease relates to effects of ARTS1 on peptide presentation, this relationship would inform research into the mechanism underlying the association of HLA B27 with ankylosing spondylitis. Second, ARTS1 cleaves cell surface receptors for the pro inflammatory cytokines IL 1 (IL 1R2)[36], IL 6 (IL 6Rα)[37] and TNF (TNFR1)[38], thereby downregulating their signaling. Genetic variants that alter the functioning of ARTS1 could therefore have pro inflammatory effects through this mechanism.

In addition to their association with ankylosing spondylitis, polymorphisms in *IL23R* have been recently documented in Crohn's disease[26,27] and psoriasis[28], suggesting that this gene is a common susceptibility factor for the major 'seronegative' diseases, at least partially explaining their co occurrence. IL 23R is a key factor in the regulation of a newly defined effector T cell subset, T_H17 cells. T_H17 cells were originally identified as a distinct subset of T cells expressing high levels of the pro inflammatory cytokine IL 17 in response to stimulation, in addition to IL 1, IL 6, TNFα, IL 22 and IL 25 (IL 17E). IL 23 has been shown to be important in the mouse models of experimental autoimmune encephalomyelitis[39], collagen induced arthritis[40] and inflammatory bowel disease[41], but it has not been studied in ankylosing spondylitis, either in human or other animal models of the disease. These studies show that blocking IL 23 reduces inflammation in these models, suggesting that the *IL23R* variants associated with disease are pro inflammatory. Successful treatment of Crohn's disease has been reported with anti IL 12p40 antibodies, which block both IL 12 and IL 23, as these cytokines share the IL 12p40 chain[42]. No functional studies of *IL23R* variants have been reported to date, and it is unclear to what extent findings in studies targeting IL 23 can be generalized to mechanisms by which *IL23R* variation affects disease susceptibility. Our genetic findings provide notable insight into the etiopathogenesis of ankylosing spondylitis and suggest that treatments targeting IL 23 may prove effective for this condition, but clearly much more needs to be understood about the mechanism underlying the observed association.

Despite the successful identification of the *ARTS1* and *IL23R* genes, it is likely either that additional real associations are present in our data but were overlooked because of their modest effect sizes, or that our focus on non synonymous coding changes led us to miss real loci. The issue of limited statistical power is emphasized in studies of nonsynonymous coding changes, which have a greater number of rare variants than other genetic variants and thus will require even larger sample sizes unless the effect sizes are larger. Other analytical approaches, such as assessing evidence for association between clusters of rare variants rather than individual loci, may prove highly informative in this regard[43], but most of the nsSNPs available in this study exist either by themselves in each gene or with one or two others, which precludes these assessments (**Supplementary Fig. 6** online). In our analyses, *ARTS1* was the only locus showing exceptional statistical significance in the scan of 1,000 cases and 1,500 controls, thus emphasizing the need for greater statistical power. We increased power by expanding the controls, or 'reference set,' to include some or all of the other disease samples. When we did so, *ARTS1* showed even stronger association evidence, the *IL23R* SNPs increased to a level that began to delineate them from background noise, and the AITD/*TSHR* confirmation emerged. This demonstration of increased statistical power through the combination of multiple datasets is timely, given the international impetus to make genotype data available to the scientific community. Future investigations will be needed to assess the power versus confounding effects and the statistical corrections needed to combine more heterogeneous samples from broader sampling regions.

These results also highlight the question of how much information may be missed by focusing on coding SNPs rather than searching more broadly over the genome at large. This question is relevant because the tradeoff between SNP panel and sample size selection is a salient factor in the design of every genome wide study. In the HapMap data[44], a substantial portion of the common nonsynonymous variation in our nsSNP set is captured by available genome wide panels (about 65% of common (MAF > 5%) nsSNPs in the Illumina Human NS 12 Beadchip are tagged with an $r^2 > 0.8$ using the Affymetrix 500 K chip, rising to 90% in the Illumina Human Hap300, which includes almost all of the nsSNPs from the NS 12 Beadchip). The four primary associated variants flagged in our study (that is, in *ARTS1*, *IL23R*, *TSHR* and *FCRL3*) would have been detected using any of the genome wide panels, because either the markers themselves or a SNP in high LD with them ($r^2 \geq 0.78$) are present on the genome wide chips. This LD relationship also emphasizes the fact that observing an association with a nsSNP does not necessarily imply that the nsSNP is causal, as it may be indirectly associated with other genetic variants in or outside the gene. Given this high degree of overlap, the continuously increasing coverage of many available genotyping products and concomitant pressures to decrease assay costs, these data suggest that future gene centric scans will be efficiently subsumed by the more encompassing and less hypothesis driven genome wide SNP panels.

METHODS

Subjects. Individuals included in the study were self identified as white and of European ancestry and came from mainland UK (England, Scotland and Wales, but not Northern Ireland). The 1,500 control samples were from the British 1958 Birth Cohort (58C, also known as the National Child Development Study), which included all the births in England, Wales and Scotland that occurred during 1 week in 1958. Recruitment details and diagnostic criteria for each of the four case groups, as well as for the North American AS replication cohort and the 58C are further described in the **Supplementary Methods** online.

Sample quality assurance and control genome wide identity by state (IBS) sharing was calculated for each pair of individuals in the combined sample of cohorts to identify first and second degree relatives whose data might contaminate the study. One subject from any pair of individuals who shared <400 genotypes IBS 0 and/or >80% alleles IBS (that is, the individual with the most missing genotypes) was removed from all subsequent analyses. To identify individuals who might have ancestries other than Western European, we merged each of our cohorts with the 60 western European (CEU) founder, 60 Nigerian (YRI) founder, and 90 Japanese (JPT) and Han Chinese (CHB) individuals from the International HapMap Project[44]. We calculated genome wide IBD distances for each pair of individuals (that is, 1 minus average IBS sharing) on those markers shared between HapMap and our nonsynonymous panel, and then used the multidimensional scaling option in R to generate a two dimensional plot based upon individuals' scores on the first two principal coordinates from this analysis (**Supplementary Fig. 2**). Any WTCCC sample that was not present in the main cluster with the CEU individuals was excluded from subsequent analyses. Finally, any individual with >10%

of genotypes missing was removed from the analysis. The number of individuals remaining after these quality control measures were applied is shown in **Table 1**.

Genotyping. We genotyped a total of 14,436 nsSNPs across the genome on all case and control samples. Because three of the diseases were of autoimmune etiology, we also typed an additional 897 SNPs within the MHC region, as well as 103 SNPs in pigmentation genes specifically designed to differentiate between population groups. SNP genotyping was performed with the Infinium I assay (Illumina), which is based on allele specific primer extension (ASPE) and the use of a single fluorochrome. The assay requires ~250 ng of genomic DNA, which is first subjected to a round of isothermal amplification that generates a 'high complexity' representation of the genome with most loci represented at usable amounts. There are two allele specific probes (50 mers) per SNP, each on a different bead type; each bead type is present on the array an average of 30 times (and a minimum of 5 times), allowing for multiple independent measurements. We processed six samples per array. Clustering was carried out with the GenCall software version 6.2.0.4, which assigns a quality score to each locus and an individual genotype confidence score that is based on the distance of a genotype from the center of the nearest cluster. First, we removed samples with more than 50% of loci having a quality score below 0.7 and then all loci with a quality score below 0.2. After clustering, we applied two additional filtering criteria: (i) we omitted individual genotypes with a genotype confidence score <0.15 and (ii) we removed any SNP for which more than 20% of samples had genotype confidence scores <0.15. The above criteria were designed to optimize genotype accuracy and minimize uncalled genotypes.

Statistical analysis markers that were monomorphic in both case and control samples, SNPs with $>10\%$ missing genotypes and SNPs with differences in the amount of missing data between cases and controls ($P < 10^{-4}$ as assessed by χ^2 test) were excluded from all analyses involving that case group only. In addition, any marker that failed an exact test of Hardy Weinberg equilibrium in controls ($P < 10^{-7}$) was excluded from all analyses[45].

Cochran Armitage tests for trend[46] were conducted using the PLINK program[47]. For the present analyses, we used the significance thresholds of $P < 10^{-4}$ 10^{-6}, as suggested for gene based scans with stronger prior probabilities than scans of anonymous markers[21]. In the present context, the lower thresholds are similar to Bonferroni significance levels (Bonferroni corrected P 0.05 corresponds to nominal P 3×10^{-6}). The conditional logistic regression analyses involving the *LNPEP* and *ARTS1* SNPs were carried out using Purcell's WHAP program[48].

We manually rechecked the genotype calls of every nsSNP with an asymptotic significance level of $P < 10^{-3}$ by inspecting raw signal intensity values and their corresponding automated genotype calls. Notably, this flagged an additional 33 markers with clear problems in genotype calling, which were subsequently excluded from all analyses (**Supplementary Fig. 4**). These results indicate that this genotyping platform generally yields highly accurate genotypes, but errors do occur and can be distributed nonrandomly between cases and controls despite stringent quality control procedures. It is imperative to check the clustering of the most significant SNPs to ensure that evidence for associations is not a result of genotyping error.

Although great lengths were taken to ensure that our samples were as homogenous as possible in terms of genetic ancestry, even subtle population substructure can substantially influence tests of association in large genome wide analyses involving thousands of individuals[49]. We therefore calculated the genomic control inflation factor, λ (ref. 20), for each case control sample as well as in the analyses where we combined the other case groups with the control individuals (**Table 2**). In general, values for λ were small (~1.1), indicating a small degree of substructure in UK samples that induces only a slight inflation of the test statistic under the null hypothesis, consistent with the results from our companion paper[12]. We therefore present uncorrected results in all analyses reported.

Consent was granted from ethical review boards of the institutions with which the participants were affiliated, and informed consent was obtained from the individuals involved in the WTCCC. Individual level data from this study will be widely available through the Consortium's Data Access Committee (http://www.wtccc.org.uk).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank the Wellcome Trust for supporting this study, and all the individuals and controls who participated in this study.

AITD: We thank the collection coordinators, J. Carr-Smith and all contributors to the AITD national DNA collection of index cases and family members from centres including Birmingham, Bournemouth, Cambridge, Cardiff, Exeter, Leeds, Newcastle and Sheffield. Principal leads for the AITD UK national collection are S.C. Gough (Birmingham), S.H.S. Pearce (Newcastle), B. Vaidya (Exeter), J.H. Lazarus (Cardiff), A. Allahabadia (Sheffield), M. Armitage (Bournemouth), P.J. Grant (Leeds) and V.K. Chatterjee (Cambridge).

Ankylosing spondylitis: We thank the Arthritis Research Campaign (UK). MAB is funded by the National Health and Medical Research Council (Australia). TASC is funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases grants 1PO1-052915-01,RO1 AR046208 and RO1-AR048465, as well as by University of Texas at Houston CTSA grant UL1RR024148, Cedars-Sinai GCRC grant MO1-RR00425, The Rosalind Russell Center for Arthritis Research at The University of California San Francisco, and the Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, US National Institutes of Health. We thank R. Jin for technical assistance and L. Diekman, L. Guthrie, F. Lin and S. Morgan for their study coordination.

Breast cancer: The breast cancer samples were clinically and molecularly curated with the assistance of A. Renwick, A. Hall, A. Elliot, H. Jayatilake, T. Chagtai, R. Barfoot, P. Kelly and K. Spanova. Our research is supported by United States Army Medical Research and Material Command grant no. W81XWH-05-1-0204, The Institute of Cancer Research and Cancer Research UK.

MS: Our work has been supported by the Wellcome Trust (grant ref. 057097), the Medical Research Council (UK) (grant ref. G0000648), the Multiple Sclerosis Society of Great Britain and Northern Ireland (grant ref. 730/02) and the National Institutes of Health (USA) (grant ref. 049477). A.G. is a postdoctoral fellow of the Research Foundation–Flanders (FWO–Vlaanderen).

L. Galver and P. Ng at Illumina and J. Morrison at the Sanger Institute contributed in the design of the nsSNP array. We thank the DNA team of the JDRF/WT DIL and T. Dibling, C. Hind and D. Simpkin at the Sanger Institute for carrying out the genotyping. We also thank S. Bingham and the WTCCC Inflammatory Bowel Disease group for genotyping the *ARTS1* markers in their replication samples.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596 604 (2007).
2. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881 885 (2007).
3. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087 1093 (2007).
4. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).
5. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989 994 (2007).
6. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638 644 (2007).
7. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39**, 977 983 (2007).
8. Moffatt, M.F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470 473 (2007).
9. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336 1341 (2007).
10. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341 1345 (2007).
11. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331 1336 (2007).
12. WTCCC. Genome-wide association studies of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 683 (2007).
13. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.* **39**, 207 211 (2007).
14. Jorgenson, E. & Witte, J.S. Coverage and power in genomewide association studies. *Am. J. Hum. Genet.* **78**, 884 888 (2006).
15. Smyth, D.J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.* **38**, 617 619 (2006).

16. Miretti, M.M. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 634 646 (2005).
17. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243 1246 (2005).
18. Sims, A.M. *et al.* Non-B27 MHC associations of ankylosing spondylitis. *Genes Immun.* **8**, 115 123 (2007).
19. McGinnis, R., Shifman, S. & Darvasi, A. Power and efficiency of the TDT and case-control design for association scans. *Behav. Genet.* **32**, 135 144 (2002).
20. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997 1004 (1999).
21. Thomas, D.C. & Clayton, D.G. Betting odds and genetic associations. *J. Natl. Cancer Inst.* **96**, 421 423 (2004).
22. Jiang, H.R. *et al.* Sialoadhesin promotes the inflammatory response in experimental autoimmune uveoretinitis. *J. Immunol.* **177**, 2258 2264 (2006).
23. Dechairo, B.M. *et al.* Association of the TSHR gene with Graves' disease: the first disease specific locus. *Eur. J. Hum. Genet.* **13**, 1223 1230 (2005).
24. Hiratani, H. *et al.* Multiple SNPs in intron 7 of thyrotropin receptor are associated with Graves' disease. *J. Clin. Endocrinol. Metab.* **90**, 2898 2903 (2005).
25. Miettinen, O.S. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am. J. Epidemiol.* **99**, 325 332 (1974).
26. Duerr, R.H. *et al.* A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* (2006).
27. Tremelling, M. *et al.* IL23R variation determines susceptibility but not disease phenotype in inflammatory bowel disease. *Gastroenterology* **132**, 1657 1664 (2007).
28. Cargill, M. *et al.* A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am. J. Hum. Genet.* **80**, 273 290 (2007).
29. Simmonds, M.J. *et al.* Contribution of single nucleotide polymorphisms within FCRL3 and MAP3K7IP2 to the pathogenesis of Graves' disease. *J. Clin. Endocrinol. Metab.* **91**, 1056 1061 (2006).
30. Kochi, Y. *et al.* A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.* **37**, 478 485 (2005).
31. Capon, F. *et al.* Fine mapping of the PSORS4 psoriasis susceptibility region on chromosome 1q21. *J. Invest. Dermatol.* **116**, 728 730 (2001).
32. Dai, K.Z. *et al.* The T cell regulator gene SH2D2A contributes to the genetic susceptibility of multiple sclerosis. *Genes Immun.* **2**, 263 268 (2001).
33. Chang, S.C., Momburg, F., Bhutani, N. & Goldberg, A.L. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism. *Proc. Natl. Acad. Sci. USA* **102**, 17107 17112 (2005).
34. Saveanu, L. *et al.* Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* **6**, 689 697 (2005).
35. Brown, M.A. *et al.* HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom. *Ann. Rheum. Dis.* **55**, 268 270 (1996).
36. Cui, X., Rouhani, F.N., Hawari, F. & Levine, S.J. Shedding of the type II IL-1 decoy receptor requires a multifunctional aminopeptidase, aminopeptidase regulator of TNF receptor type 1 shedding. *J. Immunol.* **171**, 6814 6819 (2003).
37. Cui, X., Rouhani, F.N., Hawari, F. & Levine, S.J. An aminopeptidase, ARTS-1, is required for interleukin-6 receptor shedding. *J. Biol. Chem.* **278**, 28677 28685 (2003).
38. Cui, X. *et al.* Identification of ARTS-1 as a novel TNFR1-binding protein that promotes TNFR1 ectodomain shedding. *J. Clin. Invest.* **110**, 515 526 (2002).
39. Cua, D.J. *et al.* Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. *Nature* **421**, 744 748 (2003).
40. Murphy, C.A. *et al.* Divergent pro- and antiinflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation. *J. Exp. Med.* **198**, 1951 1957 (2003).
41. Hue, S. *et al.* Interleukin-23 drives innate and T cell-mediated intestinal inflammation. *J. Exp. Med.* **203**, 2473 2483 (2006).
42. Mannon, P.J. *et al.* Anti-interleukin-12 antibody for active Crohn's disease. *N. Engl. J. Med.* **351**, 2069 2079 (2004).
43. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869 872 (2004).
44. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299 1320 (2005).
45. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887 893 (2005).
46. Armitage, P. Test for linear trend in proportions and frequencies. *Biometrics* **11**, 375 386 (1955).
47. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559 575 (2007).
48. Purcell, S., Daly, M.J. & Sham, P.C. WHAP: haplotype-based association analysis. *Bioinformatics* **23**, 255 256 (2007).
49. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512 517 (2004).

The complete list of authors is as follows:

Wellcome Trust Case Control Consortium

Management Committee: Paul R Burton[1], David G Clayton[2], Lon R Cardon[3,55], Nick Craddock[4], Panos Deloukas[5], Audrey Duncanson[6], Dominic P Kwiatkowski[3,5], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Nilesh J Samani[10], John A Todd[2], Peter Donnelly (Chair)[11]
Analysis Committee: Jeffrey C Barrett[3], Paul R Burton[1], Dan Davison[11], Peter Donnelly[11], Doug Easton[12], David M Evans[3], Hin Tak Leung[2], Jonathan L Marchini[11], Andrew P Morris[3], Chris CA Spencer[11], Martin D Tobin[1], Lon R Cardon[3,55], David G Clayton[2]
UK Blood Services & University of Cambridge Controls: Antony P Attwood[5,8], James P Boorman[8,9], Barbara Cant[8], Ursula Everson[13], Judith M Hussey[14], Jennifer D Jolley[8], Alexandra S Knight[8], Kerstin Koch[8], Elizabeth Meech[15], Sarah Nutland[2], Christopher V Prowse[16], Helen E Stevens[2], Niall C Taylor[8], Graham R Walters[17], Neil M Walker[2], Nicholas A Watkins[8,9], Thilo Winzer[8], John A Todd[2], Willem H Ouwehand[8,9]
1958 Birth Cohort Controls: Richard W Jones[18], Wendy L McArdle[18], Susan M Ring[18], David P Strachan[19], Marcus Pembrey[18,20]
Bipolar Disorder (Aberdeen): Gerome Breen[21], David St Clair[21]; (Birmingham): Sian Caesar[22], Katharine Gordon Smith[22,23], Lisa Jones[22]; (Cardiff): Christine Fraser[23], Elaine K Green[23], Detelina Grozeva[23], Marian L Hamshere[23], Peter A Holmans[23], Ian R Jones[23], George Kirov[23], Valentina Moskivina[23], Ivan Nikolov[23], Michael C O'Donovan[23], Michael J Owen[23], Nick Craddock[23]; (London): David A Collier[24], Amanda Elkin[24], Anne Farmer[24], Richard Williamson[24], Peter McGuffin[24]; (Newcastle): Allan H Young[25], I Nicol Ferrier[25]
Coronary Artery Disease (Leeds): Stephen G Ball[26], Anthony J Balmforth[26], Jennifer H Barrett[26], Timothy D Bishop[26], Mark M Iles[26], Azhar Maqbool[26], Nadira Yuldasheva[26], Alistair S Hall[26]; (Leicester): Peter S Braund[10], Paul R Burton[1], Richard J Dixon[10], Massimo Mangino[10], Suzanne Stevens[10], Martin D Tobin[1], John R Thompson[1], Nilesh J Samani[10]
Crohn's Disease (Cambridge): Francesca Bredin[27], Mark Tremelling[27], Miles Parkes[27]; (Edinburgh): Hazel Drummond[28], Charles W Lees[28], Elaine R Nimmo[28], Jack Satsangi[28]; (London): Sheila A Fisher[29], Alastair Forbes[30], Cathryn M Lewis[29], Clive M Onnie[29], Natalie J Prescott[29], Jeremy Sanderson[31], Christopher G Matthew[29]; (Newcastle): Jamie Barbour[32], M Khalid Mohiuddin[32], Catherine E Todhunter[32], John C Mansfield[32]; (Oxford): Tariq Ahmad[33], Fraser R Cummings[33], Derek P Jewell[33]
Hypertension (Aberdeen): John Webster[34]; (Cambridge): Morris J Brown[35], David G Clayton[2]; (Evry, France): Mark G Lathrop[36]; (Glasgow): John Connell[37], Anna Dominiczak[37]; (Leicester): Nilesh J Samani[10]; (London): Carolina A Braga Marcano[38], Beverley Burke[38], Richard Dobson[38], Johannie Gungadoo[38], Kate L Lee[38], Patricia B Munroe[38], Stephen J Newhouse[38], Abiodun Onipinla[38], Chris Wallace[38], Mingzhan Xue[38], Mark Caulfield[38]; (Oxford): Martin Farrall[39]
Rheumatoid Arthritis: Anne Barton[40], The Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee*, Ian N Bruce[40], Hannah Donovan[40], Steve Eyre[40], Paul D Gilbert[40], Samantha L Hilder[40], Anne M Hinks[40], Sally L John[40], Catherine Potter[40], Alan J Silman[40], Deborah PM Symmons[40], Wendy Thomson[40], Jane Worthington[40]
Type 1 Diabetes: David G Clayton[2], David B Dunger[2,41], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], Barry Widmer[2,41], John A Todd[2]
Type 2 Diabetes (Exeter): Timothy M Frayling[42,43], Rachel M Freathy[42,43], Hana Lango[42,43], John R B Perry[42,43], Beverley M Shields[43], Michael N Weedon[42,43], Andrew T Hattersley[42,43]; (London): Graham A Hitman[44]; (Newcastle): Mark Walker[45]; (Oxford): Kate S Elliott[3,7], Christopher J Groves[7], Cecilia M Lindgren[3,7], Nigel W Rayner[3,7], Nicolas J Timpson[3,46], Eleftheria Zeggini[3,7], Mark I McCarthy[3,7]

Tuberculosis (Gambia): Melanie Newport[47], Giorgio Sirugo[47], (Oxford): Emily Lyons[3], Fredrik Vannberg[3], Adrian V S Hill[3]
Ankylosing Spondylitis: Linda A Bradbury[48], Claire Farrar[49], Jennifer J Pointon[49], Paul Wordsworth[49], Matthew A Brown[48,49]
Autoimmune Thyroid Disease: Jayne A Franklyn[50], Joanne M Heward[50], Matthew J Simmonds[50], Stephen CL Gough[50]
Breast Cancer: Sheila Seal[51], Breast Cancer Susceptibility Collaboration (UK)*, Michael R Stratton[51,52], Nazneen Rahman[51]
Multiple Sclerosis: Maria Ban[53], An Goris[53], Stephen J Sawcer[53], Alastair Compston[53]
Gambian Controls (Gambia): David Conway[47], Muminatou Jallow[47], Melanie Newport[47], Giorgio Sirugo[47]; (Oxford): Kirk A Rockett[3], Dominic P Kwiatkowski[3,5]
DNA, Genotyping, Data QC and Informatics (Wellcome Trust Sanger Institute, Hinxton): Suzannah J Bumpstead[5], Amy Chaney[5], Kate Downes[2,5], Mohammed JR Ghori[5], Rhian Gwilliam[5], Sarah E Hunt[5], Michael Inouye[5], Andrew Keniry[5], Emma King[5], Ralph McGinnis[5], Simon Potter[5], Rathi Ravindrarajah[5], Pamela Whittaker[5], Claire Widden[5], David Withers[5], Panos Deloukas[5]; (Cambridge): Hin Tak Leung[2], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], John A Todd[2]
Statistics (Cambridge): Doug Easton[12], David G Clayton[2]; (Leicester): Paul R Burton[1], Martin D Tobin[1]; (Oxford): Jeffrey C Barrett[3], David M Evans[3], Andrew P Morris[3], Lon R Cardon[3,55]; (Oxford): Niall J Cardin[11], Dan Davison[11], Teresa Ferreira[11], Joanne Pereira Gale[11], Ingeleif B Hallgrimsdóttir[11], Bryan N Howie[11], Jonathan L Marchini[11], Chris CA Spencer[11], Zhan Su[11], Yik Ying Teo[3,11], Damjan Vukcevic[11], Peter Donnelly[11]
Principal Investigators: David Bentley[5,54], Matthew A Brown[48,49], Lon R Cardon[3,55], Mark Caulfield[38], David G Clayton[2], Alastair Compston[53], Nick Craddock[23], Panos Deloukas[5], Peter Donnelly[11], Martin Farrall[39], Stephen CL Gough[50], Alistair S Hall[26], Andrew T Hattersley[42,43], Adrian V S Hill[3], Dominic P Kwiatkowski[3,5], Christopher G Matthew[29], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Miles Parkes[27], Marcus Pembrey[18,20], Nazneen Rahman[51], Nilesh J Samani[10], Michael R Stratton[51,52], John A Todd[2], Jane Worthington[40]
AITD Replication Group: Sarah L Mitchell[50], Paul R Newby[50], Oliver J Brand[50], Jackie Carr Smith[50], Simon H S Pearce[56], Stephen C L Gough[50]
IL23R replication: R McGinnis[5], A Keniry[5], P Deloukas[5], TASC.
The Australo Anglo American Spondylitis Consortium (TASC): John D Reveille[57], Xiaodong Zhou[57], Linda A Bradbury[58], Anne Marie Sims[58], Alison Dowling[58], Jacqueline Taylor[58], Tracy Doan[58], Lon R Cardon[55,59], John C Davis[60], Jennifer J Pointon[61], Laurie Savage[62], Michael M Ward[63], Thomas L Learch[64], Michael H Weisman[65], Paul Wordsworth[61], Matthew A Brown[58,61]

***See Supplementary Note for details.**

Affiliations for participants: [1]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. [2]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [4]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [7]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. [8]Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 2PT, UK. [9]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 2PT, UK. [10]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. [11]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [12]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [13]National Health Service Blood and Transplant, Sheffield Centre, Longley Lane, Sheffield S5 7JN, UK. [14]National Health Service Blood and Transplant, Brentwood Centre, Crescent Drive, Brentwood CM15 8DP, UK. [15]The Welsh Blood Service, Ely Valley Road, Talbot Green, Pontyclun CF72 9WB, UK. [16]The Scottish National Blood Transfusion Service, Ellen's Glen Road, Edinburgh EH17 7QT, UK. [17]National Health Service Blood and Transplant, Southampton Centre, Coxford Road, Southampton SO16 5AF, UK. [18]Avon Longitudinal Study of Parents and Children, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK. [19]Division of Community Health Services, St. George's University of London, Crammer Terrace, London SW17 0RE, UK. [20]Institute of Child Health, University College London, 30 Guilford St., London WC1N 1EH, UK. [21]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [22]Department of Psychiatry, Division of Neuroscience, Birmingham University, Birmingham B15 2QZ, UK. [23]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff Univeristy, Heath Park, Cardiff CF14 4XN, UK. [24]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. [25]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne NE1 4LP, UK. [26]LIGHT and LIMM Research Institutes, Faculty of Medicine and Health, University of Leeds, Leeds LS1 3EX, UK. [27]IBD Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 2QQ, UK. [28]Gastrointestinal Unit, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU UK. [29]Department of Medical & Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Hospital, London SE1 9RT, UK. [30]Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK. [31]Department of Gastroenterology, Guy's and St. Thomas' NHS Foundation Trust, London SE1 7EH, UK. [32]Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. [33]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK. [34]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. [35]Clinical Pharmacology Unit and the Diabetes and Inflammation Laboratory, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [36]Centre National de Genotypage, 2 Rue Gaston Cremieux, Evry, Paris 91057, France. [37]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK. [38]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London, Queen Mary's School of Medicine, Charterhouse Square, London EC1M 6BQ, UK. [39]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [40]arc Epidemiology Research Unit, University of Manchester, Stopford Building, Oxford Rd, Manchester M13 9PT, UK. [41]Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK. [42]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter EX1 2LU, UK. [43]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, Barrack Road, Exeter EX2 5DU UK. [44]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [45]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [46]The MRC Centre for Causal Analyses in Translational Epidemiology, Bristol University, Canynge Hall, Whiteladies Rd., Bristol BS2 8PR, UK. [47]MRC Laboratories, Fajara, The Gambia. [48]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Woolloongabba, Queensland 4102, Australia. [49]Botnar Research Centre, University of Oxford, Headington, Oxford OX3 7BN, UK. [50]Department of Medicine, Division of Medical Sciences, Institute of Biomedical Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [51]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. [52]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [53]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [54]Illumina Cambridge, Chesterford Research Park, Little Chesterford, NR Saffron Walden, Essex CB10 1XL, UK. [55]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [56]University of Newcastle, Institute for Human Genetics, Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. [57]Rheumatology and Clinical Immunogenetics, University of Texas Houston Medical School, Houston, Texas 77030, USA. [58]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, University of Queensland, Brisbane 4072, Australia. [59]Statistical Genetics, Wellcome Trust Centre for Human Genetics, Oxford OX37BN, UK. [60]Department of Rheumatology, University of California, San Francisco, California 94143, USA. [61]Botnar Research Centre, University of Oxford, Oxford OX37BN, UK. [62]The Spondylitis Association of America, Sherman Oaks, California 91403, USA. [63]National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [64]Department of Radiology and [65]Department of Medicine/Rheumatology, Cedars Sinai Medical Center, Los Angeles, California 90048, USA.

ARTICLES

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

Copy number variants (CNVs) account for a major proportion of human genetic polymorphism and have been predicted to have an important role in genetic susceptibility to common disease. To address this we undertook a large, direct genome-wide study of association between CNVs and eight common human diseases. Using a purpose-designed array we typed ~19,000 individuals into distinct copy-number classes at 3,432 polymorphic CNVs, including an estimated ~50% of all common CNVs larger than 500 base pairs. We identified several biological artefacts that lead to false-positive associations, including systematic CNV differences between DNAs derived from blood and cell lines. Association testing and follow-up replication analyses confirmed three loci where CNVs were associated with disease—*IRGM* for Crohn's disease, HLA for Crohn's disease, rheumatoid arthritis and type 1 diabetes, and *TSPAN8* for type 2 diabetes—although in each case the locus had previously been identified in single nucleotide polymorphism (SNP)-based studies, reflecting our observation that most common CNVs that are well-typed on our array are well tagged by SNPs and so have been indirectly explored through SNP studies. We conclude that common CNVs that can be typed on existing platforms are unlikely to contribute greatly to the genetic basis of common human diseases.

Genome wide association studies (GWAS) have been extremely suc cessful in associating SNPs with susceptibility to common diseases, but published SNP associations account for only a fraction of the genetic component of most common diseases, and there has been considerable speculation about where the 'missing heritability'[1] might lie. Chromosomal rearrangements can cause particular rare diseases and syndromes[2], and recent reports have suggested a role for rare CNVs, either individually or in aggregate, in susceptibility for a range of common diseases, notably neurodevelopmental diseases[3–6]. So far, there have been relatively few reported associations between common diseases and common CNVs (see for example refs 7–11), which might simply reflect incomplete catalogues of common CNVs or the lack of reliable assays for their large scale typing. Here we report the results of our direct association study, identify the popu lation properties of the set of CNVs studied, describe novel analytical methods to facilitate robust analyses of CNV data, and document artefacts that can afflict CNV studies.

We designed an array to measure copy number for the majority of a recently compiled inventory of CNVs from an extensive discovery experiment[12], and several other sources. We then used the array to type 3,000 common controls and 2,000 cases of each of the diseases: bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, hyper tension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. These eight diseases make a major impact on public ill health[13], cover a range of aetiologies and genetic predispositions, and have been extensively studied via SNP based GWAS, including our earlier Wellcome Trust Case Control Consortium (WTCCC) study[14].

Pilot experiment, array content, assay and samples

Pilot experiment. We undertook a pilot experiment to compare three different platforms for assaying CNVs and to assess the merits of different experimental design parameters (see Supplementary Information for full details). On the basis of the pilot data, we chose the Agilent Comparative Genomic Hybridization (CGH) platform, and aimed to target each CNV with ten distinct probes, although in the analyses below we include any CNV targeted by at least one probe (Supplementary Fig. 9). Our analysis of the pilot CGH data indicated that the quality of the copy number signal for genotyping (rather than for discovery) at a CNV is reduced when the reference sample is homozygous deleted, in effect because the reference channel then just measures noise. To minimize this effect we used a fixed pool of DNAs as the reference sample throughout our main experiment.

Array content. Informed by our pilot experiment, we designed the CNV typing array in a collaboration with the Genome Structural Variation Consortium (GSV) in which a preliminary set of candidate CNVs was shared at an early stage with the WTCCC. Table 1 sum marizes the design content of the array, and Fig. 1 illustrates the various categories of designed loci unsuitable for association analysis. (See Methods for further details.)

Assay. In brief (see Supplementary Information for further details), the Agilent assay differentially labels parallel aliquots of the test sample and reference DNA (a pool of genomic UK lymphoblastoid cell line DNAs from nine males and one female prepared in a single batch for all experiments) and then combines them, hybridizes to the array, washes and scans. Intensity measurements for the two different labels are made at each probe separately for the test and reference DNA. These act as surrogates for the amount of DNA present, with analyses typically relying on the ratio of test to reference intensity measure ments at each probe.

Samples. A total of 19,050 case control samples were sent for assay ing: ~2,000 for each of the eight diseases and ~3,000 common con trols (these were equally split between the 1958 British Birth Cohort

*Lists of authors and their affiliations appear at the end of the paper.

Table 1 | Discovery source for regions targeted on the genotyping array

Source of loci	Number of loci targeted	Number of loci analysed	Number of loci polymorphic with good calls
CNVs			
GSV discovery project	10,835	10,217	3,096
Affymetrix 500k	18	14	12
Affymetrix 6.0	83	81	47
Illumina 1M	82	81	18
WTCCC CNV loci	231	209	108
Novel sequence			
Novel insert regions	292	292	151
Total	11,541	10,894	3,432

GSV CNVs were prioritized according to extent of polymorphism in European discovery samples. See Methods for full details of other sources.

(58C) and the UK Blood Services (UKBS) controls). These were augmented by 270 HapMap1 samples (see ref. 12 for additional analyses of the HapMap data) and 610 duplicate samples for quality control purposes. About 80% of samples from the WTCCC SNP GWAS were used here. (See Supplementary Information for further details of sample collections, inclusion criteria, and so on.)

Data pre-processing, CNV calling and quality control

Data pre processing. For each sample, raw data from the CNV experiment consist of intensity measurements for the test and reference sample for each probe. There are numerous choices at the data pre processing stage, including how to normalize data to reduce inter individual variation, and how to combine the information across the set of probes within a CNV. Several novel analytical tools substantially improved data quality, but no single approach works well for every CNV, so we carried through 16 pre processing pipelines to maximize the number of CNVs that can be tested for association. (See Supplementary Information Section 4 for illustrations and a sense of the challenges.)

CNV calling. The objective in CNV calling at each CNV is to assign each assayed sample to a diploid copy number class, which represents the sum of copy numbers on each allele. This step is analogous to, but typically considerably more challenging than, calling genotypes from SNP chip data. Available assays for SNPs are more robust and have better signal to noise properties than do available assays for CNVs[15]. We used two different statistical methods ('CNVtools', which is available as a Bioconductor package, and 'CNVCALL') in parallel to estimate the number of copy number classes at each CNV and assign individuals to these classes. (See Supplementary Information for further details.) Figure 2 illustrates three multi allelic CNVs that have attracted attention in the literature in part due to the difficulties in obtaining reliable data.

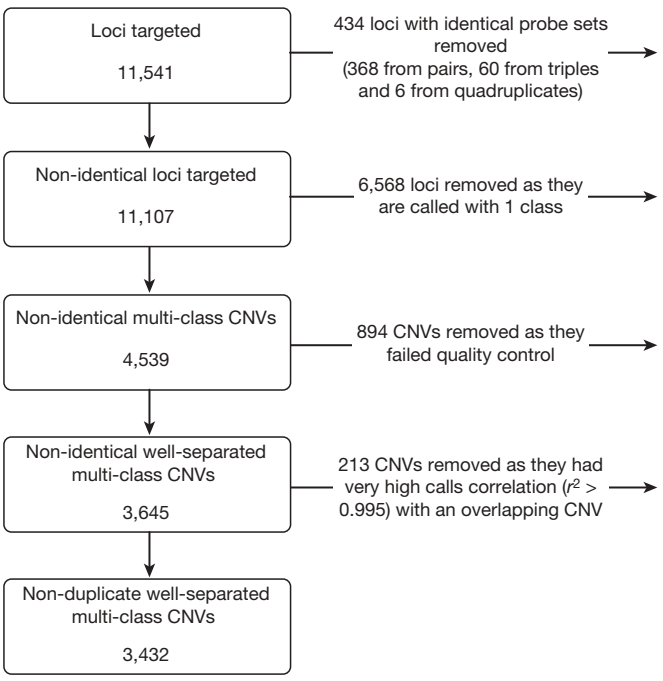


Figure 1 | Flowchart showing which CNVs are included on the array. The chart shows the reasons for CNVs being removed from consideration (the column of arrows and text to the right of the figure) from those originally targeted on the array, and the number of CNVs remaining at each stage of filtering.

Quality control. After the application of quality control metrics to each sample and each CNV (see Methods), 17,304 case control samples (of 19,050 initially) were available for association testing. There were 3,432 CNVs with more than one copy number class which passed quality control and were included in subsequent analyses. At these CNVs, concordance of calls between pairs of duplicate samples was 99.7%.

Properties of CNVs

Single class CNVs. Of the 10,894 distinct putative CNVs typed on the array after removal of detectable redundancies, 60% are called with a single copy number class, and so cannot be tested for association. After detailed analyses (see Methods) we estimate that just under half of these are probably not polymorphic. For the remainder, the combination of the experimental assay and analytical methods we have used do not allow separate copy number classes to be distinguished.

Multi class CNVs. A total of 4,326 CNVs were called with multiple classes. Of these, 3,432 passed quality control filters, which in practice means that the classes were well separated and thus that it was possible to assign individuals to copy number classes with high confidence. Most of these CNVs (88%) have two or three copy number classes, consistent with their having only two variants, or alleles, present in the population (we refer to these as bi allelic CNVs). Note that some loci involving both duplications and deletions could be called with only three classes if both homozygote classes are very rare.

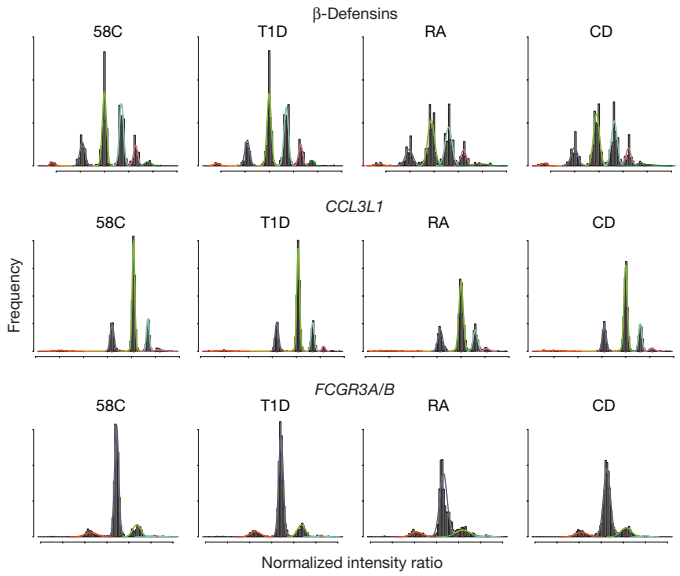


Figure 2 | Illustrative CNVs. Histograms of three multi allelic CNVs (one per row) previously reported to be associated with autoimmune diseases: β defensin (CNVR3771.10), *CCL3L1* (CNVR7077.12) and *FCGR3A/B* (CNVR383.1), showing 6, 5 and 4 fitted copy number classes, respectively. The histogram of normalized intensity ratios is shown for one control and the three autoimmune collections. Histograms are overlaid by the fitted distribution used to model each class (variously the red, blue, light green, cyan, magenta and dark green curves). In all such figures, the area under the fitted curve of a particular colour is the same for all collections at the same CNV. 58C, 1958 British Birth Cohort; CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type 1 diabetes.

Allele frequencies. Supplementary Fig. 21 shows the distribution of minor allele frequency (MAF) for bi allelic CNVs passing quality con trol. For example, 44% of autosomal CNVs passing quality control had MAF <5%. This is shifted towards lower MAFs compared to com monly used SNP chips. One consequence is that for given sample sizes association studies will tend to have lower power than for SNP studies. (See Supplementary Fig. 22 for power estimates.) Extrapolating from analyses described in ref. 12 gives an estimate that the 3,432 CNVs we directly tested represent 42 50% of common (MAF >5%) CNVs greater than 0.5 kilobases (kb) in length which are polymorphic in a population with European ancestry.

Tagging by SNPs. In the literature discussing the possible role of common CNVs in human disease there has been controversy over the extent to which CNVs will be in linkage disequilibrium with SNPs. If linkage disequilibrium between CNVs and SNPs were similar to that between SNPs, SNPs typed in GWAS would act as tags not only for untyped SNPs but also for untyped CNVs, and in turn SNP based GWAS would have indirectly explored CNVs for association with disease. (See refs 16 and 17 for opposite views.) Our large scale geno typing of an extensive CNV catalogue allows us to settle this question. In fact, CNVs that are typed well in our experiment are in general well tagged by SNPs almost to the same extent that SNPs are well tagged by SNPs (Supplementary Fig. 20). Among variable 2 and 3 class CNVs passing quality control with MAF >10%, 79% have $r^2 > 0.8$ with at least one SNP; for those with MAF <5%, 22% have $r^2 > 0.8$ with at least one SNP. This is consistent with the vast majority having arisen from unique mutational events at some time in the past. It follows that genetic variation in the form of common CNVs which can be typed on our array, has already been explored indirectly for association with common human disease through the SNP based GWAS. In passing, we note that the high correlations between our CNV calls and SNP genotypes provide strong indirect evidence that our CNV calls are capturing real variation. It is possible that the CNVs that we cannot type well are systematically different from those that we can type, for example in having many more copy number classes, and hence perhaps that they arise from repeated mutational events in the same region, in which case their linkage disequilibrium properties with SNPs could also be systematically different from the CNVs that we can type. We have no data that bear on this question, and it seems likely that such CNVs will be difficult to type genome wide on any currently available platforms.

Association testing

We performed association testing at each of the CNVs that passed quality control, in two parallel approaches. First, we applied a frequentist likelihood ratio association test that combines calling (using CNVtools) and testing into a single procedure, using an exten sion of an approach previously described[18]. Second, we undertook Bayesian association analyses in which the posterior probabilities from CNVCALL were used to calculate a Bayes factor to measure strength of association with the disease phenotypes. Important features of both sets of analyses are that they correctly handle un certainty in assignment of individuals to copy number classes, and by allowing for some systematic differences in intensities between cases and controls, that they provide robustness against certain artefacts which could arise from differences in data properties between cases and controls. There were no substantial differences between the broad conclusions from the frequentist and Bayesian approaches.

Our association analyses were based on a model in which a single parameter quantifies the increase in disease risk between successive copy number classes, analogous to that underlying the trend test for SNP data. Various analyses of the robustness of our procedure, adequacy of the model, and lack of population structure were encouraging (see Methods and Supplementary Information). For example, Supplementary Fig. 23 shows quantile quantile plots for the primary comparison of each case collection against the combined controls, and for the analogous comparisons between the two control groups. These show generally good agreement with the expectation under the null hypothesis.

Careful analysis of our association testing revealed several sophi sticated biological artefacts that can lead to false positive associa tions. These include dispersed duplications, whereby the variation at a CNV is not in the chromosomal location in the reference sequence to which the probes in the CNV uniquely match, and a DNA source effect whereby particular CNVs, and genome wide intensity data, can look systematically different according to whether the assayed DNA was derived from blood or cell lines. (See Box 1 for illustrations and further details.)

Independent replication of putative association signals is a routine and essential aspect of SNP based association studies. Particularly in view of the differences in data quality between SNP assays and CNV assays, and the wide range of possible artefacts in CNV studies, rep lication is even more important in the CNV context. Several possible approaches to replication are available. When a CNV is well tagged by a SNP (or SNPs), replication can be undertaken by assessment of the signal at the tag SNP(s) in an independent sample, either by typing the SNP or by reference to published data. Where no SNP tag is available, direct typing of the CNV in independent samples is necessary, either using a qualitative breakpoint assay or a quantitative DNA dosage assay. In most cases there will be a choice of assays. Notably, replica tion via SNPs was possible for 15 out of 18 of the CNVs for which we undertook replication based on analysis of our penultimate data freeze.

Figure 3 plots *P* values for the primary frequentist analysis for each CNV in each collection. Table 2 provides details of the top, replicated, association signals in our experiment after visual inspection of cluster plots to detect artefacts not removed by earlier quality control. Cluster plots for each CNV in Table 2 are shown in Supplementary Figs 18 and 19, and Supplementary Files 2 and 3.

There is one positive control for the diseases we studied, namely the known CNV association at the *IRGM* locus in Crohn's disease[7]. Reassuringly, our study found this association (P 1×10^{-7}, odds ratio (OR) 0.68; throughout, all ORs are with respect to increasing copy number).

We identified three loci HLA for Crohn's disease, rheumatoid arthritis and type 1 diabetes; *IRGM* for Crohn's disease; and *TSPAN8* for type 2 diabetes at which CNVs seemed to be associated with disease, all of which we convincingly replicated through previously typed SNPs that tag the CNV, and a fourth locus (CNV7113.6) at which there is suggestive evidence for association and replication in both Crohn's disease and type 1 diabetes.

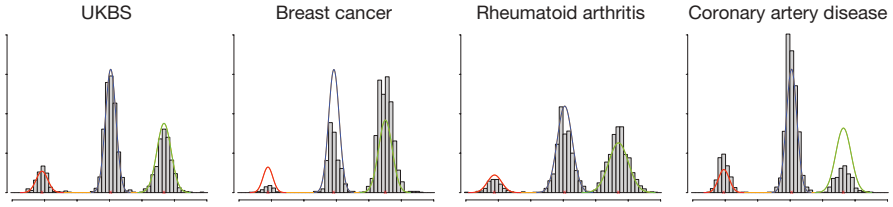
We observed CNVs in the HLA region associated variously with Crohn's disease (CNVR2841.20, P 1.2×10^{-5}, OR 0.80), rheumatoid arthritis (CNVR2845.14, P 1.4×10^{-39}, OR 1.77) and type 1 diabetes (CNVR2845.46, P 8×10^{-153}, OR 0.2). Copy number variation has previously been documented on various HLA haplotypes[19] and owing to the extensive linkage disequilibrium in the region it is perhaps not unexpected to have found CNV asso ciations in our direct study. Linkage disequilibrium across the HLA region has hampered attempts to fine map causal variation across this locus, and we have no evidence that suggests that the HLA CNVs associated with autoimmune diseases in this study represent signals independent of the known associated haplotypes.

We identified two distinct CNVs 22 kb apart upstream of the *IRGM* gene, both of which are associated with Crohn's disease. The longer CNV (CNVR2647.1, P 1.0×10^{-7}, OR 0.68) has previ ously been identified[7] as a possible causal variant on an associated haplotype first identified through SNP GWAS[14], and acted as our positive control; however, the association of the smaller CNV (CNVR2646.1, P 1.1×10^{-7}, OR 0.68, located <2 kb down stream from a different gene, *C5orf62*) is a novel observation. Although direct experimental evidence links the associated haplo types with variation in expression of the *IRGM* gene, it does not bear on the question of which of the two CNVs or the associated SNPs

Box 1 | Some artefacts in CNV association testing

Some types of artefacts, such as population structure and calling artefacts, are very similar to those seen in SNP studies. Others, related to differences in data properties between cases and controls, can be potentially more serious for CNVs[26,27]. In this box we draw attention to some specific artefacts of biological interest that we observed and which researchers should consider as explanations of putative disease-relevant associations. We note that, for the unwary, some of these artefacts could easily survive 'replication' of an association.

First, we consider dispersed CNVs. Box 1 Fig. 1 shows cluster plots for a particular CNV (CNVR2664.1) that shows a strong case-control association signal for breast cancer cases ($P = 5 \times 10^{-143}$, higher copy number for disease) with a related signal for rheumatoid arthritis ($P = 3 \times 10^{-27}$), and a signal in the opposite direction for coronary artery disease ($P = 4 \times 10^{-30}$). The right-hand class (green curve) has a higher frequency in breast cancer (and rheumatoid arthritis), and a lower frequency in coronary heart disease. (The area under the green curve is the same for each collection.) This turned out to be an artefact caused by differences in sex ratio in the various case and control samples (breast cancer, 100% female; rheumatoid arthritis, 74% female; coronary artery disease, 22% female; controls, 50% female). Comparing breast cancer cases against female controls abolished the signal. The CNV is annotated as being on chromosome 5 and all 10 probes in the CNV map uniquely to chromosome 5 in the human reference sequence. However, we found that SNPs which tagged the variation at this CNV all mapped to the X chromosome and that the region containing the probes for this CNV is present on the X chromosome in the Venter genome. We conclude that the CNV is a dispersed duplication, with the variation actually occurring on the X chromosome, and not on chromosome 5. We found one similar example, of a CNV (CNVR1065.1, featured in Table 2 as a replicated association) annotated as mapping uniquely to chromosome 2 that shows a strong signal in type 1 diabetes and rheumatoid arthritis. Careful examination shows it to be another dispersed duplication where the polymorphism is located in the HLA region, and is well tagged by HLA SNPs known to be associated with both diseases. Supplementary Fig. 27 shows the clear evidence from interchromosomal linkage disequilibrium that these two loci are dispersed duplications.

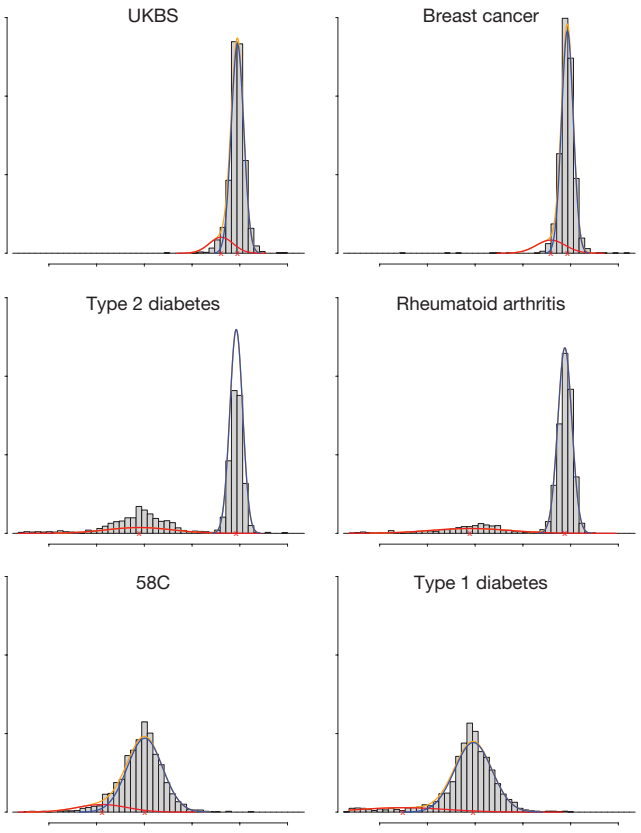


Box 1 Fig. 1 | Dispersed duplications leading to false-positive associations.

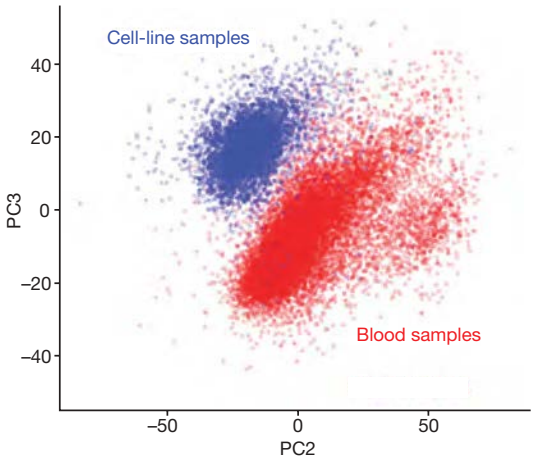
Second, we consider variation in DNA source. Box 1 Fig. 2 shows cluster plots for a different CNV (CNVR866.8) with marked differences in type 2 diabetes as compared with the UKBS controls (or against just the 58C controls). The plots show histograms of normalized intensity ratios for six collections. Examination of the pattern across collections is interesting. The collections in the top row show a single tight peak towards the right of the plot. Those in the bottom row show a single, more dispersed peak to the left. The collections in the middle row show evidence of both peaks. It turns out that for collections with the tight peak all DNA samples were derived from blood whereas all samples in the two collections with the single dispersed peak had DNA derived from cell lines. The remaining collections contain some DNAs derived from both sources. This CNV (and many others) thus exhibit systematically different behaviour depending on the DNA source. Box 1 Fig. 3 shows a plot of the second (PC2) and third (PC3) principal components of the array-wide intensity data (plot created using all samples after quality control from all ten collections using data from all CNVs, with each point representing one sample, with the points coloured according to whether that sample was derived from blood (red) or cell lines (blue)). It is clear that these two components can almost perfectly classify samples according to the source of the DNA.

Lymphoblastoid cell lines are typically grown from transformed B cells, whereas DNA extracted from blood comes largely from a mixture of white blood cells. One specific feature of B cells is that each B cell has been subject to its own pattern of rearrangements around the immunoglobulin genes via the process of V(D)J recombination[28]. This suggests a natural candidate for our observed DNA source effect, and indeed the CNV illustrated in Box 1 Fig. 2 is located close to one of the immunoglobulin genes, as are the other instances we have found of similar gross DNA source effects. But it is not the whole story. Principal components analysis of genome-wide intensity data with any probe mapping to within 1 megabase of an immunoglobulin gene excluded from analysis (Supplementary Fig. 29) shows reasonably clear discrimination by DNA source (although less clear than when all probes are included), with many probes, genome-wide, contributing to the discrimination.

Dispersed duplications and DNA source effects represent interesting biological artefacts. We also observed more prosaic effects. As one example, Supplementary Fig. 30 shows that there are systematic effects on probe intensity of the row of the plate in which a sample was run.



Box 1 Fig. 2 | DNA source effect leading to false-positive associations.



Box 1 Fig. 3 | Principal component analysis showing DNA source effect.

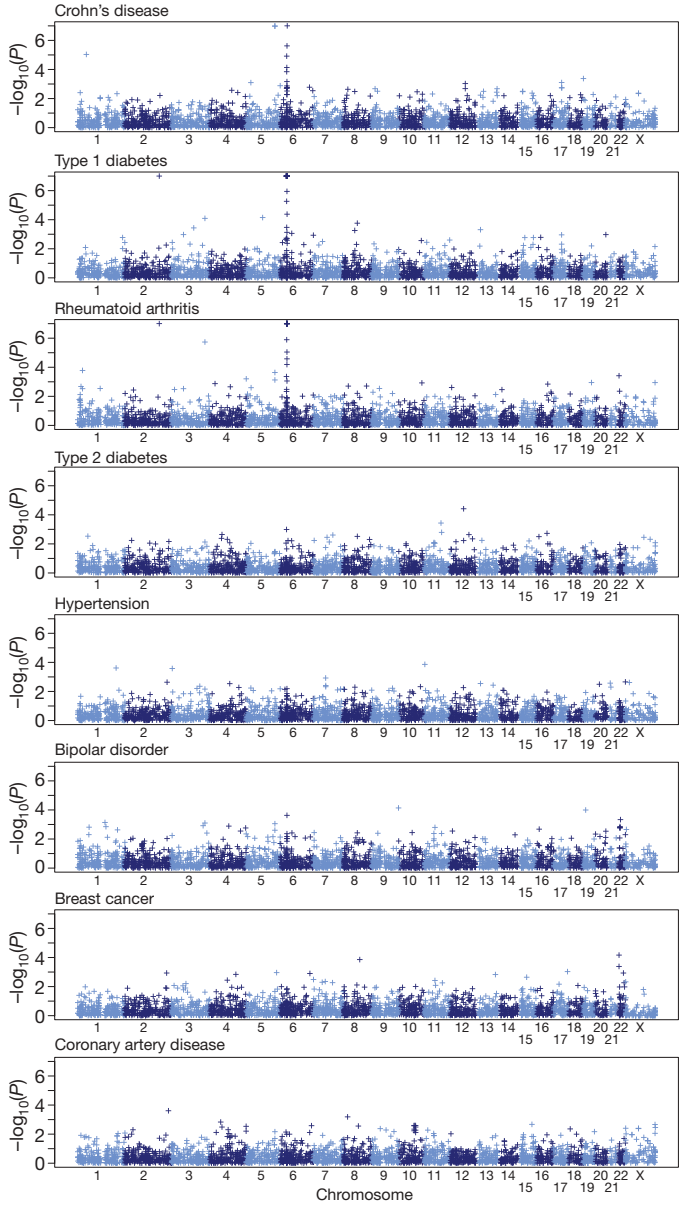


Figure 3 | Genome-wide association results. Distribution of $-\log_{10}(P)$ along the 23 chromosomes where P is the P value for the one degree of freedom test of association for each disease. The x axis shows the chromosomes numbered from 1 (on the left) to X (on the right). CNVs included in these plots were filtered on the basis of a clustering quality score (see Supplementary Information for details) and manual inspection of the most significant associations. The two apparent associations on chromosome 2 for rheumatoid arthritis and type 1 diabetes result from a dispersed duplication in which the variation is actually located within the HLA locus (see Box 1).

might be driving this variation[7]. Our conditional regression analyses on the two CNVs and SNPs on this haplotype do not point significantly to any one of these as being more strongly associated.

SNP variation in the *TSPAN8* locus was recently shown to be reproducibly associated with type 2 diabetes[20], but the potential role of a CNV is a novel observation. This CNV (CNVR5583.1, P 3.9×10^{-5}, OR 0.85) potentially encompasses part or all of an exon of *TSPAN8* and so is a plausible causal variant. The most significantly associated SNP identified in the recent meta analysis is only weakly correlated with the CNV as originally tested (r^2 0.17), and so the CNV may simply be weakly correlated with the true causal variant. Closer examination of probe level data at this CNV indicates a series of different events (including an inverted duplication and a deletion) resulting in more complex haplotypes than those tested for association by our automated approach. With this more refined definition of haplotypes the signal is stronger. (See Supplementary Information for details.)

CNVR7113.6 lies within a cluster of segmentally duplicated sequences that demarcate one end of a common 900 kb inversion polymorphism on chromosome 17 that has previously been shown to be associated with number of children and higher meiotic recombination in females[21]. The CNV shows weak evidence for association with Crohn's disease (P 1.8×10^{-3}, OR 1.15) and type 1 diabetes (P 1.1×10^{-3}, OR 1.13), but is in extremely high linkage disequilibrium (r^2 1) with SNPs known to tag the inversion, and so is in tight linkage disequilibrium with a long haplotype spanning many possible causal variants. This CNV encompasses at least one spliced transcript, but no high confidence gene annotations. Fine mapping the causal variant within such a long, tightly linked haplotype is likely to prove challenging.

In addition to the loci in Table 2, we undertook replication on 13 other loci, detailed in Supplementary Table 13, for which there was some evidence of association ($P < 1 \times 10^{-4}$ or $\log_{10}$(Bayes factor (BF)) > 2.1) in our analysis of the penultimate data freeze. Replication results were negative for all these loci. Several other loci for which there is weak evidence ($P < 1 \times 10^{-4}$ or $\log_{10}$(BF) > 2.6) for association in our final data analysis are listed in Supplementary Table 14.

To investigate further the potential role of CNVs as pathogenically relevant variants underlying published SNP associations, we took 94 association intervals in type 1 diabetes, Crohn's disease and type 2 diabetes (excluding the HLA), and for the index SNP in each association interval assessed its correlation with our calls at 3,432 CNVs. We identified two index SNPs as being correlated with an r^2 of greater than 0.5 with a called CNV. The SNPs were: rs11747270 with both CNVR2647.1 and CNVR2646.1 (*IRGM*), and rs2301436 with CNVR3164.1 (*CCR6*), both for Crohn's disease. Both of these association intervals were also identified in an independent analysis using CNV calls on HapMap samples by ref. 12.

As a further test of our approach, we examined three multiallelic CNVs that have attracted attention in the literature, both for the challenges of obtaining reliable data and for putative associations with a range of autoimmune diseases: *CCL3L1* (our CNVR7077.12); β defensins (CNVR3771.10); and *FCGR3A/B* (CNVR383.1)[10,22–24]. Encouragingly, all three CNVs pass quality control and give good quality data. Figure 2 shows cluster plots for these CNVs in our experiment. The best calls for the three CNVs required the use of two analysis pipelines (sets of choices about normalization and probe summaries) different from our standard pipeline. None of the CNVs shows significant association with the three autoimmune diseases in our study after allowance for multiple testing. In particular, we do not see formally significant evidence to replicate the reported association for *CCL3L1* and rheumatoid arthritis[24] (nominal P 0.058).

We also assessed whether CNVs that delete all or part of exons might be enriched among disease susceptibility loci, even if our study were not well powered enough to see statistically significant evidence of association for individual CNVs. To do so, we compared the 53 exonic deletion CNVs[12] that passed quality control with collections of CNVs of the same size, matched for MAF and numbers of classes. We used a (two sided) Wilcoxon signed rank test[25] to ask whether the strength of signal for association (measured by Bayes Factors) was systematically different for the exon deletion CNVs as compared to the matched CNVs. We found no evidence that deletion of an exon systematically changed evidence for association (see Supplementary Information). In a related analysis, we compared CNVs passing quality control that were well tagged by SNPs ($r^2 > 0.8$) to those passing quality control that were not, again matching for MAF and number of classes (excluding low MAF CNVs and those failing Hardy Weinberg equilibrium tests to avoid calling artefacts). There was no evidence that CNVs passing quality control that are not well tagged by SNPs are enriched for stronger signals of association compared to those which were well tagged (see Supplementary Information).

Table 2 | Replicated CNV associations and those at replicated loci

Disease	Chr.	Start (bp) (CNV)	Length (kb)	Locus	Fitted no. classes*	Combined controls (*P*)†	Extended reference‖ (*P*)	Combined controls ($\log_{10}$(BF))‡	Extended reference‖ ($\log_{10}$(BF))	Combined controls (OR)§	Extended reference‖ (OR)	MAF		Replication size		Replication size (*P*)
												Ctrls¶	Cases#	Ctrls	Cases	
T2D	12	69,818,942 (CNVR5583.1)	1.0	*TSPAN8*	3	3.9×10^{-5}	2.5×10^{-6}	2.8	4.3	0.85	0.85	0.40	0.36	5,579	4,549☆	3.9×10^{-5}
CD	5	150,157,836 (CNVR2646.1)	3.9	*IRGM*	3	1.1×10^{-7}	5.5×10^{-5}	5.8	4.1	0.68	0.75	0.07	0.10	7,977	6,894☆	7.5×10^{-11}
CD	5	150,183,562 (CNVR2647.1)	20.1	*IRGM*	3	1.0×10^{-7}	4.3×10^{-5}	6.1	3.8	0.68	0.76	0.07	0.10	7,977	6,894☆	3.9×10^{-10}
CD	6	31,416,574 (CNVR2841.20)	5.1	HLA	3	1.7×10^{-5}	1.1×10^{-5}	3.6	3.9	0.80	0.82	0.19	0.23	NA	NA	NA
T1D	6	32,582,950 (CNVR2845.46)	6.7	HLA	2	8.0×10^{-153}	2.1×10^{-196}	125.5	154.4	0.20	0.26	0.14	0.01	NA	NA	NA
RA	6	32,609,209 (CNVR2845.14)	4.0	HLA	4	1.4×10^{-39}	8.1×10^{-60}	51.5	73.5	1.77	1.83	NA	NA	NA	NA	NA
RA	2→6	179,004,449 (CNVR1065.1)	0.8	HLA	3	6.8×10^{-49}	1.6×10^{-69}	51.0	73.7	1.85	1.94	0.36	0.49	NA	NA	NA
T1D	2→6	179,004,449 (CNVR1065.1)	0.8	HLA	3	1.3×10^{-29}	1.1×10^{-39}	28.0	38.4	1.62	1.61	0.36	0.47	NA	NA	NA
RA	NA	NA (AC_000138.1_44)	5.6	HLA	3	8.3×10^{-4}	1.1×10^{-5}	1.3	2.7	0.87	0.86	0.25	0.28	2,743	3,398	1.1×10^{-3}
T1D	NA	NA (AC_000138.1_44)	5.6	HLA	3	2.0×10^{-31}	2.7×10^{-45}	31.0	45.1	0.59	0.57	0.25	0.36	2,649	3,883	7.3×10^{-50}
CD	17	40,930,407 (CNVR7113.6)	33.9	Chr17inv	3	1.2×10^{-3}	5.8×10^{-4}	1.4	1.6	1.15	1.14	0.24	0.21	6,069	4,978☆	8.6×10^{-5}
T1D	17	40,930,407 (CNVR7113.6)	33.9	Chr17inv	3	1.6×10^{-3}	7.5×10^{-4}	1.0	1.2	1.13	1.12	0.24	0.21	9,395	7,911☆	4.6×10^{-6}

Only one of the several associated CNVs mapping to the HLA in the reference sequence is shown for each of rheumatoid arthritis, type 1 diabetes and Crohn's disease. Further details of replication assays and methods are given in Supplementary Information. AC_000138.1_44 is a novel sequence insertion present in the Venter genome sequence but not in the reference sequence and hence no chromosomal location is presented. Minor allele frequency is only estimated for CNVs with three or fewer copy number classes. CD, Crohn's disease; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes.

* The number of diploid copy number classes.

† *P* value from the frequentist association test combining UKBS and 58C as controls.

‡ The $\log_{10}$ of the Bayes factor from the Bayesian association analysis combining UKBS and 58C as controls.

§ The odds ratio estimated for each additional copy of the CNV based on both UKBS and 58C as controls.

‖ Extended reference refers to the analogous quantities calculated in comparing cases of the disease in question with UKBS, 58C and aetiologically unrelated cases.

¶ The minor allele frequency in controls (UKBS plus 58C).

The minor allele frequency in cases.

☆ Replication sample includes WTCCC samples.

Discussion

We have undertaken a genome wide association study of common copy number variation in eight diseases by developing a novel array targeting most of a recently discovered set of CNVs. Our findings inform understanding of the genetic contributions to common disease, offer methodological insights into CNV analysis, and provide a resource for human genetics research.

One major conclusion is that considerable care is needed in analysing copy number data from array CGH experiments. Choices of normalization, probe summary and probe weighting can make major differences to data quality and utility in association testing. Notably, the optimal choices vary greatly across the CNVs we studied.

A second major conclusion is that CNV association analyses are susceptible to a range of artefacts that can lead to false positive associations. Some are a consequence of the less robust nature of the data compared to SNP chips. But others, such as systematic differences depending on DNA source (for example, blood versus cell lines) and dispersed duplications, are more subtle. Several artefacts could survive replication studies. Simultaneously studying eight diseases helped greatly in identifying these artefacts, and stringent quality control was invaluable in eliminating false positive associations. At least for currently available CNV typing platforms, we recommend considerable care in interpreting putative CNV associations combined with independent replication on a different experimental platform.

Despite the important technical challenges and potential artefacts discussed above, we have demonstrated that high confidence CNV calls can be assigned in large, real world case control samples for a substantial proportion of the common CNVs estimated to be present in the human genome. We have identified directly several CNV loci that are associated with common disease. Such loci could contribute to disease pathogenesis. However, the loci identified are well tagged by SNPs and, hence, the associations can be, and were, detected indirectly via SNP association studies.

There is a marked difference between the number of confirmed, replicated associations from our CNV study (3 loci) and that from the comparably sized WTCCC1 SNP GWAS of seven diseases and its immediate follow up (~24 loci). (In assessing the importance of CNVs in disease, it is the absolute number of associations, rather than the proportion among loci tested, that is important.) Following ref. 12 we estimated that our study directly tests approximately half of all autosomal CNVs >500 bp long, with MAF >5%. For such CNVs, our power averages over 80% for effects with odds ratios >1.4, and ~50% for odds ratio 1.25 (Supplementary Fig. 22). We conclude that at least for the eight diseases studied, and probably more generally, there are unlikely to be many associated CNVs with effects of this magnitude.

Might there be many more common disease associated CNVs each of small effect, in the way that we now know to be the case with SNP associations for many diseases? The total number of CNVs over 500 bp with MAF >5% is limited (estimated to be under 4,000 (ref. 12)), so unless many of these simultaneously affect many different diseases (something for which we saw no evidence outside of the HLA region) there would seem to be insufficient such CNVs for hundreds to be associated with each of many common diseases. In addition, most common CNVs (MAF >5%) are well tagged by SNPs, and thus amenable to indirect study by SNP GWAS. Examining the large meta analyses of SNP GWAS for Crohn's disease, type 1 diabetes and type 2 diabetes, there were 95 published associated loci of which only 3, including HLA, had the property that CNVs correlated with the associated SNPs; two of these were detected in our direct study.

We conclude that common CNVs typable on current platforms are unlikely to have a major role in the genetic basis of common diseases, either through particular CNVs having moderate or large effects (odds ratios >1.3, say) or through many such CNVs having small effects. In particular, such common CNVs seem unlikely to account for a substantial proportion of the 'missing heritability' for these diseases. Among the CNVs that we could type well, those not well

tagged by SNPs have the same overall association properties as those which are well tagged. We saw no enrichment of association signals among CNVs involving exonic deletions.

We have argued elsewhere[14] that the concept of 'genome wide significance' is misguided, and that under frequentist and Bayesian approaches it is not the number of tests performed but rather the prior probability of association at each locus that should determine appropriate *P* value thresholds. Here, to reduce the possibility of missing genuine associations, we deliberately set relaxed thresholds for taking CNVs into replication studies. Having completed these analyses the hypothesis that, a priori, an arbitrary common CNV is much more likely than an arbitrary common SNP to affect disease susceptibility is not supported by our data.

Limitations. Our findings should be interpreted within the context of several limitations. First, despite our successes in robustly testing some of the previously noted challenging CNVs in the genome, for some CNVs we could not reliably assign copy number classes from our assay. We estimate that just under half of these were not polymorphic in our data, being either false positives in the discovery experiment, or very rare in the UK population. For the remainder, we were also unable to perform reliable association analyses based directly on intensity measurements (that is, without first assigning individuals to copy number classes; data not shown). Such CNVs might plausibly be systematically different from those that we do type successfully, in which case it is not possible to extrapolate from our results to their potential role in human disease. Second, we note that we have not studied CNVs of sequences not present in the reference assembly, high copy number repeats such as LINE elements, or most polymorphic tandem repeat arrays, and our findings may not generalize to such variation. Finally, our experiment was powered to detect associations with common copy number variation and our observations and conclusions do not necessarily generalize to the study of rare copy number variants. Different approaches will be necessary to investigate the contribution of such variation to common disease.

METHODS SUMMARY

Pilot study. A total of 384 samples spanning a range of DNA quality were assayed for 156 previously identified CNVs on each of three different platforms: Agilent CGH, NimbleGen CGH and Illumina iSelect. The pilot experiment contained many more probes per CNV than we anticipated using in the main study, and replicates of these probes, to allow an assessment of data quality as a function of the number of probes per CNV and of the merits of replicating probes predicted in advance to perform well, compared to using distinct probes.

Sample selection. Case samples came from previously established UK collections. Control samples came from two sources: half from the 1958 Birth Cohort and half from a UK Blood Service sample. Approximately 80% of samples had been included within the WTCCC SNP GWAS study. The 610 duplicate samples were drawn from all collections.

Array design. The main study used an Agilent CGH array comprising 105,072 long oligonucleotide probes. Probes were selected to target CNVs identified mainly through the GSV discovery experiment[12], with some coming from other sources. Ten non polymorphic regions of the X chromosome were assayed for control purposes.

Array processing. Arrays were run at Oxford Gene Technology (OGT). The samples were processed in batches of 47 samples drawn from two different collections, with each batch containing one control sample for quality control purposes. These batches were randomized to protect against systematic biases in data characteristics between collections.

Data analysis. Primary data and low level summary statistics were produced at OGT. All substantive data analyses were undertaken within the consortium. Plates failing quality control metrics were rerun, as were 1,709 of the least well performing samples. Details of the common CNVs assayed in this study, including any tag SNP, are given at http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 October 2009; accepted 5 March 2010.

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747 753 (2009).
2. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10,** 451 481 (2009).
3. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445 449 (2007).
4. Stankiewicz, P. & Beaudet, A. L. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr. Opin. Genet. Dev.* **17,** 182 192 (2007).
5. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232 236 (2008).
6. The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237 241 (2008).
7. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genet.* **40,** 1107 1112 (2008).
8. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25 34 (2009).
9. de Cid, R. *et al.* Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nature Genet.* **41,** 211 215 (2009).
10. Hollox, E. J. *et al.* Psoriasis is associated with increased β defensin genomic copy number. *Nature Genet.* **40,** 23 25 (2008).
11. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459,** 987 991 (2009).
12. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature*. doi:10.1038/nature08516 (7 October 2009).
13. Murray, C. J. & Lopez, A. D. Evidence based health policy lessons from the Global Burden of Disease Study. *Science* **274,** 740 743 (1996).
14. The Wellcome Trust Case Control Consortium. Genome wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661 678 (2007).
15. McCarroll, S. A. & Altshuler, D. M. Copy number variation and association studies of human disease. *Nature Genet.* **39,** S37 S42 (2007).
16. Locke, D. P. *et al.* Linkage disequilibrium and heritability of copy number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79,** 275 290 (2006).
17. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38,** 86 92 (2006).
18. Barnes, C. *et al.* A robust statistical method for case control association testing with copy number variation. *Nature Genet.* **40,** 1245 1252 (2008).
19. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60,** 1 18 (2008).
20. Zeggini, E. *et al.* Meta analysis of genome wide association data and large scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638 645 (2008).
21. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37,** 129 137 (2005).
22. Fanciulli, M. *et al. FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ specific, autoimmunity. *Nature Genet.* **39,** 721 723 (2007).
23. Mamtani, M. *et al. CCL3L1* gene containing segmental duplications and polymorphisms in *CCR5* affect risk of systemic lupus erythaematosus. *Ann. Rheum. Dis.* **67,** 1076 1083 (2008).
24. McKinney, C. *et al.* Evidence for an influence of chemokine ligand 3 like 1 (*CCL3L1*) gene copy number on susceptibility to rheumatoid arthritis. *Ann. Rheum. Dis.* **67,** 409 413 (2008).
25. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1,** 80 83 (1945).
26. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large scale, case control association study. *Nature Genet.* **37,** 1243 1246 (2005).
27. Field, S. F. *et al.* Experimental aspects of copy number variant assays at CCL3L1. *Nature Med.* **15,** 1115 1117 (2009).
28. Lieber, M. R., Yu, K. & Raghavan, S. C. Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations. *DNA Repair* **5,** 1234 1245 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The principal funder of this project was the Wellcome Trust. Many individuals, groups, consortia, organizations and funding bodies have made important contributions to sample collections and coordination of the scientific analyses. Details are provided in Supplementary Information Section 11. We are indebted to all those who participated within the sample collections.

Author Contributions are listed in Supplementary Information.

Author Information Summary information for the CNVs studied, including genomic locations, numbers of classes and SNP tags on different platforms is available at http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml. Full data are available, under a data access mechanism, from the European Genome phenome Archive (http://www.ebi.ac.uk/ega/page.php). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.D. (peter.donnelly@well.ox.ac.uk).

The Wellcome Trust Case Control Consortium

Nick Craddock[1]*, Matthew E. Hurles[2]*, Niall Cardin[3], Richard D. Pearson[4], Vincent Plagnol[5], Samuel Robson[2], Damjan Vukcevic[4], Chris Barnes[2], Donald F. Conrad[2], Eleni Giannoulatou[3], Chris Holmes[3], Jonathan L. Marchini[3], Kathy Stirrups[2], Martin D. Tobin[6], Louise V. Wain[6], Chris Yau[3], Jan Aerts[2], Tariq Ahmad[7], T. Daniel Andrews[2], Hazel Arbury[2], Anthony Attwood[2,8,9], Adam Auton[3], Stephen G. Ball[10], Anthony J. Balmforth[10], Jeffrey C. Barrett[2], Inês Barroso[2], Anne Barton[11], Amanda J. Bennett[12], Sanjeev Bhaskar[2], Katarzyna Blaszczyk[13], John Bowes[11], Oliver J. Brand[14], Peter S. Braund[15], Francesca Bredin[16], Gerome Breen[17,18], Morris J. Brown[19], Ian N. Bruce[11], Jaswinder Bull[20], Oliver S. Burren[5], John Burton[2], Jake Byrnes[4], Sian Caesar[21], Chris M. Clee[2], Alison J. Coffey[2], John M. C. Connell[22], Jason D. Cooper[5], Anna F. Dominiczak[22], Kate Downes[5], Hazel E. Drummond[23], Darshna Dudakia[20], Andrew Dunham[2], Bernadette Ebbs[20], Diana Eccles[24], Sarah Edkins[2], Cathryn Edwards[25], Anna Elliot[20], Paul Emery[26], David M. Evans[27], Gareth Evans[28], Steve Eyre[11], Anne Farmer[18], I. Nicol Ferrier[29], Lars Feuk[30,31], Tomas Fitzgerald[2], Edward Flynn[11], Alistair Forbes[32], Liz Forty[1], Jayne A. Franklyn[14,33], Rachel M. Freathy[34], Polly Gibbs[20], Paul Gilbert[11], Omer Gokumen[35], Katherine Gordon Smith[1,21], Emma Gray[2], Elaine Green[1], Chris J. Groves[12], Detelina Grozeva[1], Rhian Gwilliam[2], Anita Hall[20], Naomi Hammond[2], Matt Hardy[5], Pile Harrison[36], Neelam Hassanali[12], Husam Hebaishi[2], Sarah Hines[20], Anne Hinks[11], Graham A Hitman[37], Lynne Hocking[38], Eleanor Howard[2], Philip Howard[39], Joanna M. M. Howson[5], Debbie Hughes[20], Sarah Hunt[2], John D. Isaacs[40], Mahim Jain[4], Derek P. Jewell[41], Toby Johnson[39], Jennifer D. Jolley[8,9], Ian R. Jones[1], Lisa A. Jones[21], George Kirov[1], Cordelia F. Langford[2], Hana Lango Allen[34], G. Mark Lathrop[42], James Lee[16], Kate L. Lee[39], Charlie Lees[23], Kevin Lewis[2], Cecilia M. Lindgren[4,12], Meeta Maisuria Armer[5], Julian Maller[4], John Mansfield[43], Paul Martin[11], Dunecan C. O. Massey[16], Wendy L. McArdle[44], Peter McGuffin[18], Kirsten E. McLay[2], Alex Mentzer[45], Michael L. Mimmack[2], Ann E. Morgan[46], Andrew P. Morris[4], Craig Mowat[47], Simon Myers[3], William Newman[28], Elaine R. Nimmo[23], Michael C. O'Donovan[1], Abiodun Onipinla[39], Ifejinelo Onyiah[2], Nigel R. Ovington[5], Michael J. Owen[1], Kimmo Palin[2], Kirstie Parnell[34], David Pernet[20], John R. B. Perry[34], Anne Phillips[47], Dalila Pinto[30], Natalie J. Prescott[13], Inga Prokopenko[4,12], Michael A. Quail[2], Suzanne Rafelt[15], Nigel W. Rayner[4,12], Richard Redon[2,48], David M. Reid[38], Anthony Renwick[20], Susan M. Ring[44], Neil Robertson[4,12], Ellie Russell[1], David St Clair[17], Jennifer G. Sambrook[8,9], Jeremy D. Sanderson[45], Helen Schuilenburg[5], Carol E. Scott[2], Richard Scott[20], Sheila Seal[20], Sue Shaw Hawkins[39], Beverley M. Shields[34], Matthew J. Simmonds[14], Debbie J. Smyth[5], Elilan Somaskantharajah[2], Katarina Spanova[20], Sophia Steer[49], Jonathan Stephens[8,9], Helen E. Stevens[5], Millicent A. Stone[50,51], Zhan Su[3], Deborah P. M. Symmons[11], John R. Thompson[6], Wendy Thomson[11], Mary E. Travers[12], Clare Turnbull[20], Armand Valsesia[2], Mark Walker[52], Neil M. Walker[5], Chris Wallace[5], Margaret Warren Perry[20], Nicholas A. Watkins[8,9], John Webster[53], Michael N. Weedon[34], Anthony G. Wilson[54], Matthew Woodburn[5], B. Paul Wordsworth[55], Allan H. Young[29,56], Eleftheria Zeggini[2,4], Nigel P. Carter[2], Timothy M. Frayling[34], Charles Lee[35], Gil McVean[3], Patricia B. Munroe[39], Aarno Palotie[2], Stephen J. Sawcer[57], Stephen W. Scherer[30,58], David P. Strachan[59], Chris Tyler Smith[2], Matthew A. Brown[55,60], Paul R. Burton[6], Mark J. Caulfield[39], Alastair Compston[57], Martin Farrall[61], Stephen C. L. Gough[14,33], Alistair S. Hall[10], Andrew T. Hattersley[34,62], Adrian V. S. Hill[4], Christopher G. Mathew[13], Marcus Pembrey[63], Jack Satsangi[23], Michael R. Stratton[2,20], Jane Worthington[11], Panos Deloukas[2], Audrey Duncanson[64], Dominic P. Kwiatkowski[2,4], Mark I. McCarthy[4,12,65], Willem H. Ouwehand[2,8,9], Miles Parkes[16], Nazneen Rahman[20], John A. Todd[5], Nilesh J. Samani[15,66] & Peter Donnelly[4,3]

*These authors contributed equally to this work.

[1]MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [3]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [4]The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [5]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. [6]Departments of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester LE1 7RH, UK. [7]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, Exeter EX1 2LU, UK. [8]Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 0PT, UK. [9]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 0PT, UK. [10]Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds LS2 9JT, UK. [11]arc Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester M13 9PT, UK. [12]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. [13]Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London SE1 9RT, UK. [14]Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research, University of Birmingham, Birmingham B15 2TT, UK. [15]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. [16]IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [17]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [18]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. [19]Clinical Pharmacology Unit, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [20]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. [21]Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham B15 2FG, UK. [22]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK. [23]Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [24]Academic Unit of Genetic Medicine, University of Southampton, Southampton SO16 5YA, UK. [25]Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK. [26]Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. [27]MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol BS8 2BN, UK. [28]Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 0JH, UK. [29]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne NE1 4LP, UK. [30]The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS Centre East Tower, 101 College St, Room 14 701, Toronto, Ontario M5G 1L7, Canada. [31]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala 75185, Sweden. [32]Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK. [33]University Hospital Birmingham NHS Foundation Trust, Birmingham B15 2TT, UK. [34]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK. [35]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [36]University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford OX3 7LD, UK. [37]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [38]Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen AB25 2ZD, UK. [39]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK. [40]Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [41]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK. [42]Centre National de Genotypage, 2 Rue Gaston Cremieux, Evry, Paris 91057, France. [43]Department of Gastroenterology and Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. [44]ALSPAC Laboratory, Department of Social Medicine, University of Bristol, Bristol BS8 2BN, UK. [45]Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK. [46]NIHR Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK. [47]Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY, UK. [48]INSERM UMR915, L'Institut du Thorax, Nantes 44035, France. [49]Clinical and Academic Rheumatology, Kings College Hospital National Health Service Foundation Trust, Denmark Hill, London SE5 9RS, UK. [50]University of Toronto, St Michael's Hospital, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada. [51]University of Bath, Claverdon, Norwood House, Room 5.11a, Bath, Somerset BA2 7AY, UK. [52]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [53]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. [54]School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield S10 2JF, UK. [55]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Windmill Road, Headington, Oxford OX3 7LD, UK. [56]UBC Institute of Mental Health, 430 5950 University Boulevard Vancouver, British Columbia V6T 1Z3, Canada. [57]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [58]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [59]Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK. [60]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland 4102, Australia. [61]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [62]Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter EX2 5DW, UK. [63]Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK. [64]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [65]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LJ, UK. [66]Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester LE3 9QP, UK.

METHODS

Pilot experiment. Full details of Methods are given in the Supplementary Information, but in brief a total of 384 samples from four different collections spanning the range of DNA quality encountered in our previous WTCCC SNP based association study[14] were assayed for 156 previously identified CNVs on each of three different platforms: Agilent Comparative Genomic Hybridization (CGH), and NimbleGen CGH (run in service laboratories) and Illumina iSelect (run at the Sanger Institute). The pilot experiment contained many more probes per CNV (40 90 depending on platform) than we anticipated using in the main study, and replicates of these probes, to allow an assessment of data quality as a function of the number of probes per CNV and of the merits of replicating probes predicted in advance to perform well, compared to using distinct probes.

The Agilent CGH platform performed best in our pilot and we settled on an array that comprised 105,072 long oligonucleotide probes. On the basis of the pilot data we aimed to target each CNV with 10 distinct probes. Actual numbers of probes per CNV on the array varied from this for several reasons (see Supplementary Information and Supplementary Fig. 9), and we included in our analyses any CNV with at least one probe on the array.

Array content, assay and samples for the main experiment. Array content: the GSV discovery experiment[12] involved 20 HapMap Utah residents with Euro pean ancestry (CEU) and 20 HapMap Yoruban (YRI) individuals, and 1 Polymorphism Discovery Resource individual, assayed via 20 NimbleGen arrays containing a total of 42,000,000 probes tiled across the assayable portion of the human reference genome. We prioritized CNVs for our experiment based on their frequency in the discovery sample, with those identified in CEU indi viduals given precedence. A total of 10,835 out of 11,700 CNVs were included from the list provided by the GSV, with those not included on the array being present in discovery in only 1 YRI individual and not overlapping genes or highly conserved elements. This list was augmented by any common CNVs not present among the GSV list found from analyses of Affymetrix SNP 6.0 data in HapMap 2 samples (83 CNVs), Illumina 1M data in HapMap 3 samples (82 CNVs), analyses of Affymetrix 500K samples (18 CNVs)[7,29,30], and from our own analyses of WTCCC1 SNP data (231 CNVs). In addition, we sought to identify CNVs not present in the human reference sequence through analyses of published[31,32] novel sequence insertions (292 CNVs in total). Thus in total, our array targeted 11,541 putative CNVs. Ten non polymorphic regions of the X chromosome were also assayed for control purposes.

Most loci targeted on the CNV typing array derive from microarray based CNV discovery, which is inherently imprecise. In contrast to SNP discovery by sequencing, arrays do not provide nucleotide level resolution, nor do they locate additional copies of a sequence in the genome. As a result, when CNVs called in different individuals overlap, but are not identical, these could be called as one or two different CNVs, and where discovered CNVs involve probes which map to multiple places in the reference genome, they might be called as CNVs in each of these locations. Interpretation of counts of CNVs from discovery experiments is thus not straightforward. Data on CNVs across thousands of individuals provide added power to refine CNV definitions and derive a non redundant set of CNVs. In addition, our CNV typing array draws together CNVs from different sources, and additional redundancy between these, although minimized during array design, can be identified and removed. Analyses of the final array design revealed 434 of the 11,541 CNVs to be redundant because they were targeted by exactly the same probes as other CNVs on the array, and analysis of our array data revealed a further 213 of 562 CNVs to be redundant from instances where overlapping CNVs passing quality control were called as distinct in discovery yet had effec tively identical copy number calls. See Supplementary Information Section 3.1 for further details on array content.

Assay: arrays were run at Oxford Gene Technology (OGT), with each plate containing one control sample for quality control purposes. Primary data and low level summary statistics were produced at OGT. All substantive data ana lyses were undertaken within the consortium. Plates that failed pre specified quality control metrics were rerun on the array, and in addition we repeated 1,709 of the least well performing samples, chosen according to our own quality control analyses. (See Supplementary Information for further details.)

Samples: the WTCCC CNV study analysed cases from eight common diseases (breast cancer, bipolar disorder, coronary artery disease, Crohn's disease, hyper tension, rheumatoid arthritis, type I diabetes, and type 2 diabetes) and two control cohorts (1958 Birth Cohort (58C) and the UK Blood Service collection (UKBS)). The number of subjects from each cohort that were analysed and the numbers that passed each phase of the quality control procedures within this study are shown in Supplementary Table 7. For bipolar disorder, coronary heart disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, type 2 diabetes and the two control cohorts, a large proportion of the subjects studied in this experi ment were the same as those in the WTCCC1 SNP GWAS (Supplementary Table 2). Where sufficient DNA was not available for the original WTCCC1 individuals, additional new samples from the same cohorts were used, selected using the same approaches used for the WTCCC1 samples. Any samples that failed any of the relevant quality control metrics in WTCCC1 were excluded from consideration for this experiment. The breast cancer cohort was not included in the WTCCC1 SNP GWAS.

Data pre processing, CNV calling and quality control. Data pre processing: for each of the targeted loci, the subset of probes that target the locus of interest (at least 1 bp overlap) while also targeting the least number of additional CNVs was selected for assaying (see Supplementary Information Section 4.2 for more details). A total of 16 different analysis 'pipelines' were used to create one dimensional intensity summaries for each CNV. First, a range of different methods were used to create single intensity measurements for each probe from the red channel (test DNA) and green channel (reference DNA) intensity data. This included different methods for normalization of the signals (see Supplementary Information Section 4.3 for details). Second, some pipelines incorporated a new method called probe variance scaling (PVS) that weights probes based on informa tion derived from duplicate samples (see Supplementary Information Section 4.5 for details). Third, some pipelines used the first principal component of the nor malized probe intensities to summarize the probe level data to CNV level data, whereas other pipelines used the mean of the probe intensities. Finally, some pipelines additionally used a linear discriminant function (LDF) to refine further the summaries based on information from an initial round of genotype calling (see Supplementary Information Section 4.4 for details).

CNV calling: algorithmic details of the two calling methods used (CNVtools and CNVCALL) are provided in Supplementary Information Section 6. Each method was applied separately to the intensity summaries created from each of the 16 pre processing pipelines for each CNV locus.

Quality control: samples were excluded on the basis of sample handling errors, evidence of non European ancestry, evidence of sample mixing, manufacturer's recommendations on data quality, outlying values of various pre calling and post calling quality metrics, and identity or close relatedness to other samples (see Supplementary Information Section 5.1 for further details). To choose which pipeline to use for a given CNV we used the pipeline that gave the highest number of classes and the highest average posterior probability in cases where more than one pipeline gave the same maximum number of classes. CNVs were excluded that had identical probe sets to other CNVs, that were called with one class in all pre processing pipelines, that had low average posterior calls in all pre processing pipelines, or that had a high calls correlation with an overlapping CNV (see Supplementary Information Section 5.2 for further details).

Properties of CNVs. Single class CNVs: Supplementary Table 15 shows the proportion of the single class CNVs from the GSV discovery project broken down according to the number of individuals and population(s) in which they were discovered. Of the GSV CNVs discovered in CEU, 52% are single class in our data, whereas a higher proportion (74%) of GSV CNVs discovered exclusively in YRI are single class, as would be expected. CNVs at which distinct copy number classes cannot be distinguished might result because: (1) although polymorphic, the signal to noise ratio at that CNV does not allow reliable identification of distinct copy number classes; (2) the copy number variant has an extremely low popu lation frequency; or (3) the CNV was a false positive in a discovery experiment and is in fact monomorphic. In a genuinely polymorphic CNV, the intensity measure ments within a pair of duplicates should be more similar than between a random pair of distinct individuals. Intensity comparisons between duplicates and random pairs of individuals thus allow estimates of the proportion of single class CNVs which are not copy number variable in our data (see Supplementary Information). These estimates range from ~23% for single class CNVs dis covered in two or more CEU individuals to ~50% of single class CNVs dis covered exclusively in YRI (see Supplementary Information for details). We estimate that 18% of GSV CNVs discovered in CEU do not exhibit polymorphism in our UK sample. This figure is similar to the GSV estimate for false positives in discovery of 15%[12]. Overall, considering CNVs on the array from all sources, we estimate that 26% do not exhibit polymorphism, so that just under half of the CNVs that seem in our data to have a single class are likely not to be polymorphic. Not all of these will be false positives in discovery; some represent CNVs that are either unique to the individual in which they were discovered or are extremely rare in the UK population.

Multi class CNVs: a companion study[12] estimated that 83% of the bi allelic CNVs it genotyped represent deletions, with the remainder being duplications. Supplementary Table 7 compares the number of copy number classes estimated by the two calling algorithms used in the analyses for each of the CNVs passing quality control. Most differences in numbers of called classes between the algo rithms arise from CNVs where one class is very rare and is handled differently by the algorithms (for example, called as a separate class in one algorithm but classed as outlier samples or merged with a larger class by the other).

These 3,432 CNVs include 80% of the CNVs genotyped on the Affymetrix 6.0 array that are common (MAF >5%) in a population with European ancestry[33]; conversely only 15% of the common CNVs we called could be called using the Affymetrix 6.0 array.

Allele frequencies: we calculated minor allele frequencies (MAFs) for 2 and 3 class CNVs by assuming that these CNVs were biallelic and using the expected posterior genotype counts (see Supplementary Information Section 7.3 for further details).

Tagging by SNPs: to determine how well tagged the CNVs analysed in our experiment were by SNPs, we carried out correlation analyses using control samples that were common to the current studies and other WTCCC studies. We analysed three different collections of SNPs. We used imputed HapMap2 SNP calls in the WTCCC1 study that used the Affymetrix 500k array, and actual calls from the WTCCC2 study using both the Affymetrix 6.0 array and a custom Illumina 1.2M array. In all cases we used samples from the UKBS collection (see Supplementary Information Section 7.1 for further details).

Geographical variation: geographical information, at the level of 13 pre defined regions of the UK, was available for 82% of the samples in our study and we undertook analyses for differences in copy number class frequencies between regions. The results, shown in Supplementary Fig. 24, confirm that there is no major genome wide population structure, but that, unsurprisingly, there is dif ferentiation at CNVs within HLA. It does not seem easy to determine whether other regions with low *P* values in this test represent genuine departures from the null hypothesis of no differentiation, rather than chance effects, although we note that the third most regionally differentiated CNV outside the HLA (CNVR7722.1, $P \quad 3 \times 10^{-5}$, 12 d.f.) is a deletion located within the gene *LILRA3*, which may act as soluble receptor for class I MHC antigens, and so would be consistent with the observed HLA stratification. This deletion is also the subject of a reported disease association[34] in multiple sclerosis, a finding that may require some caution given the level of geographical stratification at this CNV in our data. (See Supplementary Information Section 9.1 for further details.)

Association testing. Diagnostic plots such as quantile quantile and cluster plots were created using R. Cluster plots were visually inspected for all CNVs with putative associations.

Principal component analysis (PCA) was applied to the summarized intensity levels for all CNVs, and for all samples that passed quality control. Plots of the first ten principal components were coloured by various sample parameters and these revealed some of the artefacts described in Box 1.

Where possible, replication was carried out by using data from other studies for SNPs that tag the CNVs of interest. Where there was no SNP tag available, breakpoint or direct quantitative CNV assays were designed (see Supplementary Information Section 9 for further details).

We used a two sided Wilcoxon signed rank test to test for differences between distributions of Bayes factors between different subsets of CNVs (those that delete all or part of an exon versus those that do not, and CNVs that are well tagged by SNPs versus those that are not well tagged). (See Supplementary Information Section 9.5 for further details.)

Testing for population stratification: all our samples are from within the UK, and we have excluded any for which the genetic data suggest evidence of non European ancestry. All collections in this study, apart from breast cancer, were involved in the WTCCC SNP GWAS, and across these collections 80% of sam ples coincided between the two studies. Analysis of the WTCCC SNP data[14] established that population structure was not a major factor confounding asso ciation testing. Similar analyses using SNP data available for the breast cancer samples yielded similar results (data not shown). These SNP results reinforce the evidence from the quantile quantile plots in Supplementary Fig. 23 and our geographical analyses of the CNV data.

Expanded reference group analysis: in addition to our primary case control analyses, following ref. 14 we also undertook expanded reference group analyses, in which copy number class frequencies in cases for a particular disease are compared with those for controls and the other diseases with no aetiological or known genetic connection (see Supplementary Table 10 for details).

Other analyses. We used information on variability between duplicate samples to determine whether CNVs called with one class show signals of polymorphism (details are given in Supplementary Information Section 9.2).

We used estimates of the number of common autosomal CNVs segregating in a population of European ancestry from ref. 12 to estimate the coverage of common autosomal CNVs in our study (see Supplementary Information Section 9.3 for further details).

We designed a series of PCR primers to analyse further the complex signals associated with CNVR5583.1 found in the *TSPAN8* region. (See Supplementary Information Section 9.4 for further details.)

29. International HapMap Project. ⟨http://hapmap.ncbi.nlm.nih.gov/⟩ (2010).
30. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444 454 (2006).
31. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
32. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56 64 (2008).
33. McCarroll, S. A. *et al.* Integrated detection and population genetic analysis of SNPs and copy number variation. *Nature Genet.* **40,** 1166 1174 (2008).
34. Koch, S. *et al.* Association of multiple sclerosis with ILT6 deficiency. *Genes Immun.* **6,** 445 447 (2005).

Breast Cancer Res Treat (2010) 124:283 288
DOI 10.1007/s10549 010 0949 1

BRIEF REPORT

Mutation and association analysis of *GEN1* in breast cancer susceptibility

Clare Turnbull · Sarah Hines · Anthony Renwick · Deborah Hughes · David Pernet · Anna Elliott · Sheila Seal · Margaret Warren-Perry · D. Gareth Evans · Diana Eccles · Breast Cancer Susceptibility Collaboration (UK) · Michael R. Stratton · Nazneen Rahman

Received: 5 May 2010 / Accepted: 11 May 2010 / Published online: 30 May 2010

Abstract GEN1 was recently identified as a key Holliday junction resolvase involved in homologous recombination. Somatic truncating *GEN1* mutations have been reported in two breast cancers. Together these data led to the proposition that *GEN1* is a breast cancer predisposition gene. In this article we have formally investigated this hypothesis. We performed full-gene mutational analysis of *GEN1* in 176 *BRCA1/2*-negative familial breast cancer samples and 159 controls. We genotyped six SNPs tagging the 30 common variants in the transcribed region of *GEN1* in 3,750 breast cancer cases and 4,907 controls. Mutation analysis revealed one truncating variant, c.2515 2519del-AAGTT, which was present in 4% of cases and 4% of controls. We identified control individuals homozygous for the deletion, demonstrating that the last 69 amino acids of GEN1 are dispensable for its function. We identified 17 other variants, but their frequency did not significantly differ between cases and controls. Analysis of 3,750 breast cancer cases and 4,907 controls demonstrated no evidence of significant association with breast cancer for six SNPs tagging the 30 common *GEN1* variants. These data indicate that although it also plays a key role in double-strand DNA break repair, *GEN1* does not make an appreciable contribution to breast cancer susceptibility by acting as a high- or intermediate-penetrance breast cancer predisposition gene like *BRCA1*, *BRCA2*, *CHEK2*, *ATM*, *BRIP1* and *PALB2* and that common *GEN1* variants do not act as low-penetrance susceptibility alleles analogous to SNPs in *FGFR2*. Furthermore, our analyses demonstrate the importance of undertaking appropriate genetic investigations, typically full gene screening in cases and controls together with large-scale case control association analyses, to evaluate the contribution of genes to cancer susceptibility.

Keywords Breast cancer · Genetic susceptibility · DNA repair · Cancer genes

Electronic supplementary material The online version of this article (doi:10.1007/s10549 010 0949 1) contains supplementary material, which is available to authorized users.

The first two authors contributed equally.

The patients participating in this research were recruited through The Breast Cancer Susceptibility Collaboration UK (BCSC). The clinicians and counsellors making up the BCSC are listed in the Appendix

C. Turnbull · S. Hines · A. Renwick · D. Hughes · D. Pernet · A. Elliott · S. Seal · M. Warren Perry · M. R. Stratton · N. Rahman (✉)
Section of Cancer Genetics, the Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK
e mail: nazneen.rahman@icr.ac.uk

D. Gareth Evans
Regional Genetic Service, St Mary's Hospital, Manchester, UK

D. Eccles
Wessex Clinical Genetics Service, Princess Ann Hospital, Southampton, UK

M. R. Stratton
Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Introduction

Breast cancer is twice as common in women with an affected first degree relative and germline mutations in the known breast cancer predisposition genes account for <30% of this excess familial risk. Inactivating mutations in *BRCA1* and *BRCA2* are high-penetrance breast cancer

susceptibility alleles accounting for ~16%, whilst mutations in the functionally related DNA repair genes *CHEK2*, *PALB2*, *ATM* and *BRIP1* are of intermediate penetrance, and account for <3% [1 7]. Genome-wide association studies have identified 18 common variants which have been classed as low-penetrance breast cancer predisposition alleles. When combined these SNPs account for approximately 8% of familial disease risk [8 15].

The breast cancer predisposition genes *BRCA1*, *BRCA2*, *CHEK2*, *ATM*, *BRIP1* and *PALB2* are involved in double-strand break repair via the homologous recombination pathway [16 18]. This pathway repairs breaks caused by ionising radiation and mutagenic chemicals by utilising the homologous chromosome as a template for repair [19]. During this process a covalent link between each pair of homologous chromatids is formed and is known as a Holliday junction. Once repair is complete, these junctions are resolved by symmetrical nicking of the DNA strands, followed by separation and ligation to form two separate duplex molecules. Holliday junction resolvases mediate the transition from four covalently bonded chromatids to two separate duplex chromosomes [20]. In 2008, *GEN1* was identified as the gene encoding a human Holliday junction resolvase with a key a role in this recombinational repair pathway [21].

Through genome-wide sequencing of the exome in breast cancer cell lines and primary tumours, two somatic frameshift mutations in *GEN1* were identified [22, 23]. This, together with recognition of the role of GEN1 in DNA repair, led to the conclusion that constitutional *GEN1* mutations would confer susceptibility to breast cancer in a fashion analogous to some other DNA repair genes [21]. However, to date, no data to support this conclusion have been published. In order to investigate formally the contribution of *GEN1* to breast cancer susceptibility, we have undertaken mutational analysis of the full gene in 192 breast cancer cases and 184 controls and an association analysis of common variants in the vicinity of *GEN1* in constitutional DNA from 3,750 breast cancer cases and 4,907 controls.

Materials and methods

Samples

Cases were unrelated individuals with breast cancer and a family history of breast cancer that were recruited through cancer genetics clinics in the UK, through the Genetics of Familial Breast Cancer Study. Informed consent was obtained from all family members and the research was approved by the London Multicentre Research Ethics Committee (MREC/01/2/18). Samples from non-Caucasian UK ethnic groups were excluded. The extent of family history was quantified using a Family History Score, defined by the number of relatives with breast cancer, weighted by their degree of relatedness to the index case. A score of 1.0 is assigned to the index case, with an additional 0.5 for each affected 1st degree relative, and an additional 0.25 for each affected 2nd degree relative. Where an individual has bilateral cancer their score is doubled. In the *GEN1* mutation screen we utilised 192 samples with a median Family History Score of 2.75 (range was 1.75 4). All cases were negative for *BRCA1* and *BRCA2* mutations and large deletions/duplications. In the *GEN1* association study we utilised 3,750 cases with a median Family History Score of 1.75 (range 1 5.25). *BRCA1/2* mutations had either been excluded (3,304) or the status was unknown (446).

Controls were obtained from the 1958 Birth Cohort Collection, an ongoing follow-up of persons born in Great Britain in 1 week in 1958 [24]. Informed consent has been obtained for all blood samples in this collection to be used as a genetic resource. Additional controls for the genome-wide association study were obtained from the United Kingdom Blood Services Collection of Common Controls established for the Wellcome Trust Case Control study, a collection of DNA samples from consenting blood donors of the English, Scottish and Welsh Blood Services [25]. Individuals of self-reported white ethnicity and representative of gender and each geographical region were selected.

GEN1 mutation analyses

We screened genomic DNA from the familial breast cancer case and control samples through the 13 coding exons and intron exon boundaries of *GEN1* (Q17RS7) in 15 PCR fragments (Supplementary Table 1). Following PCR, we carried out uni-directional sequencing using BigDyeTerminator Cycle sequencing kit and 3730XL automated sequencer (ABI). All variants and mutations were confirmed by separate bi-directional sequencing in a different aliquot of native DNA. We analysed the coding sequence and ten intronic flanking bases of each exon using Mutation Surveyor software version 3.20 (SoftGenetics) and visual inspection. Only samples successfully analysed through at least 90% of the *GEN1* coding sequence were included; we successfully mutationally screened 176/192 cases and 159/184 controls. We assessed the likely pathogenicity of variants using Polyphen (http://genetics.bwh.harvard.edu/pph/), SIFT (http://blocks.fhcrc.org/sift/SIFT.html) and NNSplice (http://fruitfly.org:9005/seq tools/splice.html) in silico software.

To further evaluate the *GEN1* truncating variant c.2515 519delAAGTT, we extended mutation analysis of exon 13 to 536 cases and 525 controls in total. To

investigate whether the variant causes nonsense-mediated RNA decay, we extracted RNA from EBV transformed lymphoblastoid cell lines from two cases and two controls heterozygous for the variant. We used SuperScript II Reverse Transcriptase (Invitrogen) to generate cDNA which was amplified, sequenced and analysed as described above using primers Forward-AAGGAGACCAGCTGCTT CAA and Reverse-GGAAGAGGGCTATCCAAACA.

Statistical analyses

We performed comparisons of the frequencies between cases and controls of variants detected through mutational screening using a two-sided Fisher's exact test. We carried out a genome-wide association study for breast cancer susceptibility alleles genotyping 3,960 breast cancer cases on a custom Illumina Infinium 670k array. Genotype frequencies were compared with those obtained on 5,069 controls genotyped on an Illumina Infinium 1.2M array, utilising data on 594,375 SNPs that were successfully genotyped on both arrays. We excluded closely related individuals (IBS probability >0.86), individuals with >15% non-European ancestry (by computing IBS scores between participants and individuals in HapMap and using multi-dimensional scaling) and restricted analyses to individuals that were called on >97% of successfully genotyped SNPs. After these exclusions, 3,750 cases and 4,907 controls were used in the final analysis [9].

The transcribed region of *GEN1* extends from 17,798,661 to 17,830,113 bp on chromosome 2 (http://genome.ucsc.edu/) and contains 30 single nucleotide polymorphisms of minor allele frequency >0.05 (http://hapmap.ncbi.nlm.nih.gov/). Linkage disequilibrium (LD) in the region was evaluated in 90 HapMap CEU individuals using a sliding window of 1,000 kB and 10,000 SNPs. These LD data were used to select six SNPs from our dataset which tag these 30 SNPs in *GEN1* at $r^2 > 0.8$ (Supplementary Table 2). We undertook association testing using a 1 df Cochran Armitage test and a general 2 df χ^2 test. Analyses were performed using Stata10 (State College, TX, USA) and PLINK (v1.06) software [26].

Results

We successfully analysed the full coding sequence and intron exon boundaries of *GEN1* in 176 individuals with familial breast cancer and 159 controls (Table 1). We identified one truncating variant, c.2515 2519delAAGTT, a five base pair deletion in the final exon of the coding sequence. We extended the analysis of this mutation which demonstrated that it was present in similar frequencies in case and control chromosomes (47/1,072 cases vs. 47/1,050 controls) and both cases and controls homozygous for the deletion were identified (Fig. 1a c). This mutation is predicted to cause protein truncation generating a product lacking 69 amino acids (~8% of the protein) from the c-terminus. The mutation is in the last exon of *GEN1* and would be anticipated to escape nonsense-mediated RNA decay [27]. This was confirmed by analysis of cDNA from cases and controls heterozygous for c.2515 2519delAAGTT, which demonstrated equal proportions of the mutant and wild-type transcripts.

We also identified four synonymous and 13 non-synonymous *GEN1* variants. 13 variants were detected at similar frequencies in cases and controls including five common variants (frequency >0.05). Two rare non-synonymous variants were found in cases but not controls and two rare non-synonymous variants were found in controls but not cases. None of the variants were predicted to affect splicing. Only one variant, c.2692C>T p.R898C, was predicted to be deleterious by both Polyphen and SIFT algorithms but the difference in frequency between case (3/372) and control (6/360) chromosomes was not significant ($P = 0.3$) (Table 1).

We compared the frequency between 3,750 familial breast cancer cases and 4,907 controls of six SNPs which tag the 30 common variants in the genomic region encompassing *GEN1* (Supplementary Table 2). There was no evidence of significant association for any of these tag SNPs with breast cancer (Table 2).

Discussion

GEN1 was recently identified as a Holliday junction resolvase with a key role in repair of DNA double-strand breaks. This function, together with the report of somatic *GEN1* mutations in two breast cancers, led to the proposition that *GEN1* would act as a breast cancer susceptibility gene, similar to some other DNA repair genes [1 5, 7]. In these recognised breast cancer susceptibility genes, *BRCA1*, *BRCA2*, *CHEK2*, *ATM*, *BRIP1* and *PALB2*, inactivating, primarily truncating, mutations confer high or intermediate risks of breast cancer. We identified a single *GEN1* truncating mutation, c.2515 2519delAAGTT. However, this deletion was present at equal frequency in cases and controls, indicating that it is not associated with appreciable increased risk of breast cancer. The deletion is in the final exon of the gene, results in truncation of <10% of the protein, and mutant transcripts are not subjected to nonsense-mediated decay. Moreover, we identified several control individuals homozygous for the deletion, demonstrating that the last 69 amino acids of the GEN1 protein are dispensable for its function. This is consistent first with findings of Ip et al. [21] who reported that a truncated form of GEN1, lacking the C-terminal, is sufficient for Holliday junction resolvase activity and secondly with phylogenetic evidence

Table 1 Coding *GEN1* variants in breast cancer cases and controls

Variant	dbSNP[a]	Allele frequencies[b]		*P* value for association[c]
		Cases	Controls	
c.274T>A; p.S92T	rs1812152	195/358	226/346	0.1
c.428A>G; p.N143S	rs16981869	22/366	21/364	0.9
c.566G>A; p.S189N		6/362	6/328	0.9
c.607A>G; p.I203V	rs10177628	3/362	0/328	0.1
c.905G>A; p.R302H		2/382	1/344	0.6
c.988G>A; p.E330K		0/380	1/340	0.3
c.1341A>G; p.A447A	rs16983864	4/362	1/356	0.2
c.1526C>G; p.S509W		1/372	3/358	0.3
c.1638T>A; p.S546S		6/372	5/358	0.8
c.1971A>G; p.E657E	rs300168	189/384	189/350	0.5
c.2039C>T; p.T680I	rs300169	233/384	228/350	0.6
c.2445C>T; p.Y815Y		3/382	1/360	0.3
c.2449A>G; p.T817A		0/382	1/360	0.3
c.2515 2519delAAGTT		47/1072	47/1050	0.9
c.2567C>T; p.S856F		1/382	0/360	0.3
c.2619T>G; p.S873R	rs57936182	4/372	1/360	0.2
c.2644A>G; p.K882E		6/372	7/360	0.7
c.2692C>T; p.R898C	rs17315702	3/372	6/360	0.3

[a] www.ncbi.nlm.nih.gov/projects/SNP

[b] The denominator for each variant indicates the number of chromosomes successfully sequenced

[c] *P* value for two sided Fisher's exact test (1 df)

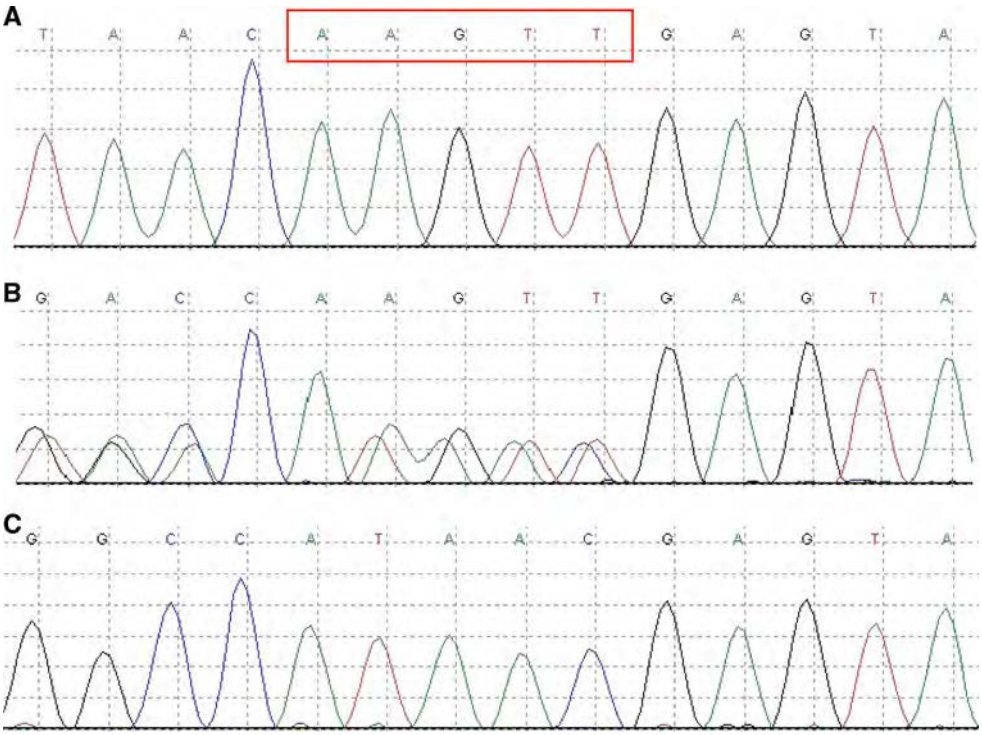


Fig. 1 Sequence traces for wild type deletion heterozygote and deletion homozygotes. Reverse sequencing chromatograms of the sequence encompassing the c.2515 2519delAAGTT deletion showing the wild type sequence (**a**) deletion heterozygote sequence (**b**) and deletion homozygote sequence (**c**). The five deleted bases are indicated by the red square in wild type sequence

which demonstrates strong conservation between GEN1 and its yeast homologue yen1 over the first 480 amino acids, but very little in the C-terminal regions [21]. Our mutation screen did not identify any additional truncating mutations, and there was no evidence that non-truncating variants are likely to be pathogenic.

Table 2 Association with breast cancer of six SNPs tagging common *GEN1* variants

Illumina tag SNP	Minor allele	MAF cases	MAF controls	*P* value[a]
rs7556886	T	0.19	0.19	0.97
rs6761104	A	0.10	0.10	0.49
rs300168	A	0.46	0.47	0.69
rs300169	G	0.36	0.36	0.42
rs17315736	A	0.09	0.09	0.49
rs13031876	C	0.35	0.35	0.60

MAF minor allele frequency

[a] Cochran Armitage trend test (1 df), unadjusted for multiple testing

Within recent years, common variants conferring small risks of breast cancer have been identified using large case control series via genome-wide analyses of single nucleotide polymorphisms [8 15, 28]. Of the 18 common, low-penetrance breast cancer susceptibility alleles identified to date, none have been in regions containing DNA repair genes. We evaluated 30 common SNPs in the vicinity of *GEN1* by comparing the frequencies of six tag SNPs in 3,750 breast cancer cases and 4,907 controls and found no evidence to suggest that any common variants in this region are associated with breast cancer.

Our mutational screening data indicate that *GEN1* does not make an appreciable contribution to breast cancer predisposition by acting as a high-penetrance breast cancer predisposition gene akin to *BRCA1* and *BRCA2* or intermediate-penetrance breast cancer predisposition gene, similar to *ATM*, *BRIP1*, *CHEK2*, or *PALB2*. The association analysis finds no evidence that common variation targeting *GEN1* confers susceptibility to breast cancer. Overall, these data strongly suggest that constitutional *GEN1* variation does not contribute to breast cancer predisposition. In addition, our analyses demonstrate the importance of undertaking appropriate genetic investigations, typically full gene screening in cases and controls together with large-scale case control association analyses, to evaluate the contribution of genes to cancer susceptibility.

Acknowledgements We thank all the patients and families that participated in this research. We thank Anita Hall and Darshna Dudakia for assistance in recruitment and Katrina Spanova and Bernadette Ebbs for running the sequencers. This work was funded by Cancer Research UK (C8620 A8372); US Military ACQ Activity, Era of Hope Award (W81XWH 05 1 0204) and the Institute of Cancer Research (UK). We acknowledge NHS funding to the NIHR Biomedical Research Centre. This study makes use of data generated by the Wellcome Trust Case Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

Appendix

The authors gratefully acknowledge the clinicians and counsellors from the Breast Cancer Susceptibility Collaboration UK (BCSC) who coordinated recruitment and collection of the FBCS samples: A. Ardern-Jones, G. Attard, K. Bailey, C. Bardsley, J. Barwell, L. Baxter, R. Belk, J. Berg, N. Bradshaw, A. Brady, S. Brant, C. Brewer, G. Brice, G. Bromilow, C. Brooks, A. Bruce, B. Bulman, L. Burgess, J. Campbell, B. Castle, R. Cetnarskyj, C. Chapman, C. Chu, N. Coates, A. Collins, J. Cook, S. Coulson, G. Crawford, D. Cruger, C. Cummings, R. Davidson, L. Day, L. de Silva, B. Dell, C. Dolling, A. Donaldson, A. Donaldson, H. Dorkins, F. Douglas, S. Downing, S. Drummond, J. Dunlop, S. Durrell, D. Eccles, C. Eddy, M. Edwards, E. Edwards, J. Edwardson, R. Eeles, F. Elmslie, G. Evans, B. Gibbens, C. Giblin, S. Gibson, S. Goff, S. Goodman, D. Goudie, L. Greenhalgh, J. Greer, H. Gregory, R. Hardy, C. Hartigan, T. Heaton, C. Higgins, S. Hodgson, T. Homfray, D. Horrigan, C. Houghton, L. Hughes, V. Hunt, L. Irvine, L. Izatt, L. Jackson, C. Jacobs, S. James, M. James, L. Jeffers, I. Jobson, W. Jones, S. Kenwrick, C. Kightley, C. Kirk, L. Kirk, E. Kivuva, A. Kumar, F. Lalloo, N. Lambord, C. Langman, P. Leonard, S. Levene, S. Locker, P. Logan, M. Longmuir, A. Lucassen, V. Lyus, A. Magee, S. Mansour, D. McBride, E. McCann, V. McConnell, M. McEntagart, K. McDermot, L. McLeish, D. McLeod, L. Mercer, C. Mercer, Z. Miedzybrodzka, J. Miller, P. Morrison, J. Myring, J. Paterson, P. Pearson, G. Pichert, K. Platt, M. Porteous, C. Pottinger, S. Price, L. Protheroe, L. Protheroe, S. Pugh, C. Riddick, V. Roffey-Johnson, M. Rogers, S. Rose, S. Rowe, A. Schofield, G. Scott, J. Scott, A. Searle, S. Shanley, S. Sharif, J. Shaw, J. Shea-Simonds, L. Side, J. Sillibourne, K. Simon, S. Simpson, S. Slater, K. Smith, L. Snadden, J. Soloway, Y. Stait, B. Stayner, M. Steel, C. Steel, H. Stewart, D. Stirling, M. Thomas, S. Thomas, S. Tomkins, H. Turner, E. Tyler, E. Wakeling, F. Waldrup, L. Walker, L. Walker, C. Watt, S. Watts, A. Webber, C. Whyte, J. Wiggins, E. Williams, L. Winchester.

References

1. Meijers Heijboer H, van den Ouweland A, Klijn J et al (2002) Low penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 muta tions. Nat Genet 31:55 59
2. Miki Y, Swensen J, Shattuck Eidens D et al (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66 71
3. Rahman N, Seal S, Thompson D et al (2007) PALB2, which encodes a BRCA2 interacting protein, is a breast cancer sus ceptibility gene. Nat Genet 39:165 167
4. Renwick A, Thompson D, Seal S et al (2006) ATM mutations that cause ataxia telangiectasia are breast cancer susceptibility alleles. Nat Genet 38:873 875

5. Seal S, Thompson D, Renwick A et al (2006) Truncating mutations in the *Fanconi anemia* J gene BRIP1 are low penetrance breast cancer susceptibility alleles. Nat Genet 38:1239 1241
6. Stratton MR, Rahman N (2008) The emerging landscape of breast cancer susceptibility. Nat Genet 40:17 22
7. Wooster R, Neuhausen SL, Mangion J et al (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12 13. Science 265:2088 2090
8. Easton DF, Pooley KA, Dunning AM et al (2007) Genome wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087 1093
9. Turnbull C, Ahmed S, Morrison J et al (2010) Genome wide association study identifies five new breast cancer susceptibility loci. Nat Genet. doi:10.1038/ng.586
10. Ahmed S, Thomas G, Ghoussaini M et al (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41:585 590
11. Hunter DJ, Kraft P, Jacobs KB et al (2007) A genome wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870 874
12. Stacey SN, Manolescu A, Sulem P et al (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor positive breast cancer. Nat Genet 39:865 869
13. Stacey SN, Manolescu A, Sulem P et al (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor positive breast cancer. Nat Genet 40:703 706
14. Thomas G, Jacobs KB, Kraft P et al (2009) A multistage genome wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41:579 584
15. Zheng W, Long J, Gao YT et al (2009) Genome wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41:324 328
16. Turnbull C, Rahman N (2008) Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet 9:321 345
17. Venkitaraman AR (2002) Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell 108:171 182
18. Zhang F, Ma J, Wu J et al (2009) PALB2 links BRCA1 and BRCA2 in the DNA damage response. Curr Biol 19:524 529
19. Khanna KK, Jackson SP (2001) DNA double strand breaks: signaling, repair and the cancer connection. Nat Genet 27:247 254
20. Lilley DM, White MF (2001) The junction resolving enzymes. Nat Rev Mol Cell Biol 2:433 443
21. Ip SC, Rass U, Blanco MG et al (2008) Identification of Holliday junction resolvases from humans and yeast. Nature 456:357 361
22. Forbes SA, Bhamra G, Bamford S et al (2008) The catalogue of somatic mutations in cancer (COSMIC). Curr Protoc Hum Genet Chap 10:Unit 10.11
23. Wood LD, Parsons DW, Jones S et al (2007) The genomic landscapes of human breast and colorectal cancers. Science 318:1108 1113
24. Weblink: National Child Development Study (2008) http://www.cls.ioe.ac.uk/studies.asp?section=000100020004. Accessed 7 Oct 2008
25. WTCC (2007) Wellcome Trust Case Control Consortium: genome wide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls. Nature 447:661 678
26. Purcell S, Neale B, Todd Brown K et al (2007) PLINK: a tool set for whole genome association and population based linkage analyses. Am J Hum Genet 81:559 575
27. Lejeune F, Maquat LE (2005) Mechanistic links between nonsense mediated mRNA decay and pre mRNA splicing in mammalian cells. Curr Opin Cell Biol 17:309 315
28. Cox A, Dunning AM, Garcia Closas M et al (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 39:352 358

nature genetics

Genome-wide association study identifies five new breast cancer susceptibility loci

Clare Turnbull[1], Shahana Ahmed[2], Jonathan Morrison[3], David Pernet[1], Anthony Renwick[1], Mel Maranian[2], Sheila Seal[1], Maya Ghoussaini[2], Sarah Hines[1], Catherine S Healey[2], Deborah Hughes[1], Margaret Warren-Perry[1], William Tapper[4], Diana Eccles[4], D Gareth Evans[5], The Breast Cancer Susceptibility Collaboration (UK)[1,10], Maartje Hooning[6], Mieke Schutte[6], Ans van den Ouweland[7], Richard Houlston[1], Gillian Ross[8], Cordelia Langford[9], Paul D P Pharoah[2,3], Michael R Stratton[1,9], Alison M Dunning[2], Nazneen Rahman[1] & Douglas F Easton[2,3]

Breast cancer is the most common cancer in women in developed countries. To identify common breast cancer susceptibility alleles, we conducted a genome-wide association study in which 582,886 SNPs were genotyped in 3,659 cases with a family history of the disease and 4,897 controls. Promising associations were evaluated in a second stage, comprising 12,576 cases and 12,223 controls. We identified five new susceptibility loci, on chromosomes 9, 10 and 11 ($P = 4.6 \times 10^{-7}$ to $P = 3.2 \times 10^{-15}$). We also identified SNPs in the 6q25.1 (rs3757318, $P = 2.9 \times 10^{-6}$), 8q24 (rs1562430, $P = 5.8 \times 10^{-7}$) and *LSP1* (rs909116, $P = 7.3 \times 10^{-7}$) regions that showed more significant association with risk than those reported previously. Previously identified breast cancer susceptibility loci were also found to show larger effect sizes in this study of familial breast cancer cases than in previous population-based studies, consistent with polygenic susceptibility to the disease.

Genome-wide association studies (GWAS) provide a powerful approach to identify common disease alleles. Recent GWAS have identified common variants at 12 loci that are associated with an increased risk of breast cancer, and an additional locus, *CASP8* (specifically, a polymorphism resulting in a D302H substitution), has been identified through a candidate-gene association study[1–8]. However, because the risks associated with these variants are modest (per-allele odds ratios (OR) <1.3), they explain only a small fraction of the estimated twofold familial relative risk of breast cancer in first-degree relatives of affected women. Moreover, the GWAS conducted to date have been relatively small, and it is likely that many susceptibility variants have been missed due to lack of power in these studies. In an attempt to identify additional breast cancer loci, we conducted a GWAS that was substantially larger than those conducted to date.

We studied 3,960 cases of breast cancer from the UK, selected for a positive family history of breast cancer. We selected cases with a positive family history because, under a polygenic model of susceptibility, this is expected to increase the effect size and hence improve study power[9]. DNA samples from these women were genotyped using an Illumina Infinium 660k array. Case genotypes were compared with those from 5,069 controls, drawn from two UK population-based studies. After quality control exclusions, we utilized data on 582,886 SNPs in 3,659 cases and 4,897 controls (Online Methods).

Genotype frequencies in cases and controls were compared using a 1-degree-of-freedom (d.f.) Cochran-Armitage trend test (**Fig. 1**; for the quantile-quantile plot see **Supplementary Fig. 1**). There was modest evidence for inflation in the test statistic ($\lambda = 1.12$, which is equivalent to $\lambda_{1,000} = 1.03$ for a study of 1,000 cases and 1,000 controls). Adjustment for differential population structure using the first two components based on a principal-components analysis of uncorrelated SNPs reduced the inflation to $\lambda = 1.06$ (Online Methods).

We observed evidence of association for all 12 of the susceptibility loci identified through previous GWAS, using the same SNP as that previously identified or a strongly correlated SNP ($P = 0.02$ to $P = 3.6 \times 10^{-31}$; **Table 1**). Seven of these loci reached $P < 10^{-4}$, among which five have previously been evaluated in large collaborative analyses of case-control studies by the Breast Cancer Association Consortium (BCAC). The BCAC analyses involved more than 20,000 cases and 20,000 controls, providing a reliable estimate of the per-allele OR[1,5,10]. For each of these five SNPs, the per-allele OR in the current study was higher than that estimated from the population-based studies by BCAC by a factor of 1.46-fold to 1.75-fold ($P < 0.05$ for difference in OR for all SNPs except rs13281615; **Supplementary Table 1**). This enrichment is broadly consistent with the selection of cases with a family history, assuming a multiplicative polygenic model (which predicts a 1.5-fold higher excess relative risk for the associated SNP for women with

[1]Section of Cancer Genetics, The Institute of Cancer Research, Sutton, Surrey, UK. [2]Department of Oncology, University of Cambridge, Strangeways Research Laboratory, Cambridge, UK. [3]Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge, UK. [4]Academic Unit of Genetic Medicine, University of Southampton, Southampton General Hospital, Southampton, UK. [5]Department of Genetic Medicine, St. Mary's Hospital, Manchester, UK. [6]Department of Medical Oncology, Erasmus University Medical Center, Rotterdam, The Netherlands. [7]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands. [8]Breast Cancer Unit, Royal Marsden National Health Service Foundation Trust, London, UK. [9]Wellcome Trust Sanger Institute, Hinxton, UK. [10]A full list of members is provided in the **Supplementary Note**. Correspondence should be addressed to D.F.E. (douglas@srl.cam.ac.uk).

Received 15 December 2009; accepted 9 April 2010; published online 9 May 2010; doi:10.1038/ng.586

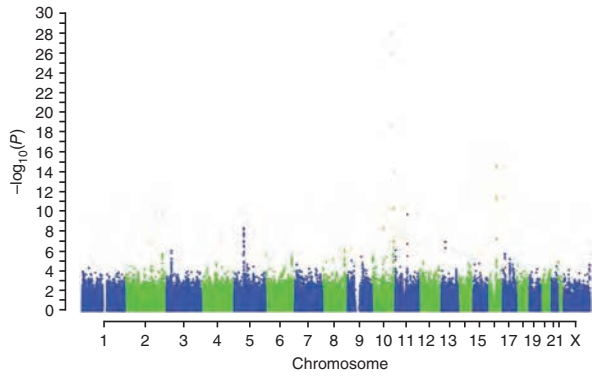


Figure 1 Manhattan plot of 1-d.f. Cochran-Armitage *P* values for association by genomic position.

one affected first-degree relative and a twofold higher excess relative risk for women with two affected first-degree relatives)[9]. The loci on 5p12 (rs7716600, a surrogate for rs10941679) and 1p11.2 do not conform to this pattern, having smaller ORs than those published previously (a 1.5-fold higher excess OR can be excluded here in each case, $P = 0.018$ and $P = 0.015$, respectively). These results suggest either that the initial effect sizes were overestimated (perhaps due to 'winner's curse') or that these loci have weaker than expected effects in women with a family history due to a different model of susceptibility than is applicable for the other loci. We also found limited evidence in support of the association with the *CASP8* D302H polymorphism ($P = 0.14$; **Table 1**)[8]. Consistent with previous results, the two loci showing the largest effect sizes and most significant associations in this GWAS were on chromosome 10, in intron 2 of *FGFR2* (rs2981579, $P = 3.6 \times 10^{-31}$) and at the *TOX3* locus on 16q (rs3803662, $P = 3.2 \times 10^{-15}$).

For three loci (6q25.1, *LSP1* and 8q24) we identified a SNP that showed a more significant association than the SNP originally reported associated to breast cancer susceptibility. The SNP with the lowest *P* value at 6q25.1 (rs3757318, $P = 2.9 \times 10^{-6}$) lies ~200 kb upstream of *ESR1* in an intron of *C6orf97*. In Europeans, rs3757318 is only weakly correlated with rs2046210, which has previously been identified as a susceptibility SNP[7] in a study from Shanghai ($r^2 = 0.088$), though these two SNPs are more strongly correlated in an East Asian population ($r^2 = 0.48$ in HapMap CHB). Both rs3757318 and rs6900157 (a surrogate for rs2046210 with $r^2 = 0.96$) remained significantly associated with breast cancer after multiple logistic regression analysis ($P = 0.0003$ and $P = 0.002$, respectively). These results suggest either the presence of a single causal variant that is more strongly correlated with rs3757318 than rs2046210 in Europeans or the presence of two causal variants. The more strongly associated SNPs that we identified in the 8q24 and *LSP1* regions lie within the same linkage disequilibrium (LD) blocks as the originally identified SNP, and in each case, the original SNP was not significantly associated with risk after adjusting for the new SNP. Thus, these results may reflect the same underlying association and should assist in narrowing the search for the true causal variants. A more strongly associated variant, rs10931936, was also identified at the *CASP8* locus ($P = 0.0014$, $r^2 = 0.13$).

After eliminating SNPs in previously identified susceptibility regions, we identified 28 SNPs in 13 regions of LD that were significant at $P < 0.00001$. After eliminating SNPs that were strongly correlated, we attempted to replicate these associations by genotyping 15 SNPs in

Table 1 Associations in the current study at previously known breast cancer loci

	Strongest association in current study				Published association				Association for published SNP in current study			
Locus	Most signifcant SNP	Alleles[a]	Per-allele OR (95% CI)[b]	*P*	Published SNP	Alleles[a]	(r^2)[c]	Published OR	Best tag in GWAS (r^2)[d]	Alleles[a]	Per-allele OR (95% CI)[b]	*P*
FGFR2	rs2981579	G/A (0.42)	1.43 (1.35–1.53)	3.6×10^{-31}	rs2981582[e]	G/A (0.38)	1.0	1.26 (1.22–1.29)[1]	rs2981579 ($r^2 = 1.0$)	G/A (0.42)	1.43 (1.35–1.53)	3.6×10^{-31}
TOX3	rs3803662	G/A (0.26)	1.30 (1.22–1.39)	3.2×10^{-15}	rs3803662	G/A (0.25)	1.0	1.19 (1.15–1.23)[1]	rs3803662	G/A (0.26)	1.30 (1.22–1.39)	3.2×10^{-15}
MAP3K1	rs889312	A/C (0.28)	1.22 (1.14–1.30)	4.6×10^{-9}	rs889312	A/C (0.38)	1.0	1.12 (1.08–1.16)[1]	rs889312	A/C (0.28)	1.22 (1.14–1.30)	4.6×10^{-9}
8q24	rs1562430	C/T (0.58)	1.17 (1.10–1.25)	5.8×10^{-7}	rs13281615	A/G (0.40)	0.42	1.08 (1.05–1.12)[1]	rs13281615	A/G (0.41)	1.14 (1.07–1.21)	2.2×10^{-5}
2q35	rs13387042	G/A (0.49)	1.21 (1.14–1.29)	2.0×10^{-10}	rs13387042	G/A (0.49)	1.0	1.12 (1.09–1.15)[10]	rs13387042	G/A (0.49)	1.21 (1.14–1.29)	2.0×10^{-10}
LSP1	rs909116	C/T (0.53)	1.17 (1.10–1.24)	7.3×10^{-7}	rs3817198	T/C (0.30)	0.23	1.07 (1.04–1.11)[1]	rs3817198	T/C (0.33)	1.12 (1.05–1.19)	0.0006
5p12	rs9790879	T/C (0.40)	1.10 (1.03–1.17)	0.0032	rs10941679	(A/G) (0.25)	0.48	1.19 (1.11–1.28)[4]	rs7716600 ($r^2 = 0.75$)	C/A (0.22)	1.11 (1.04–1.19)	0.0034
6q25.1	rs3757318	G/A (0.07)	1.30 (1.17–1.46)	2.9×10^{-6}	rs2046210	G/A (0.34)	0.088	1.15[f] (1.03–1.28)[7]	rs6900157 ($r^2 = 0.96$)	T/C (0.35)	1.15 (1.08–1.22)	1.8×10^{-5}
SLC4A7	rs4973768	C/T (0.47)	1.16 (1.10–1.24)	5.8×10^{-7}	rs4973768	C/T (0.46)	1.0	1.11 (1.08–1.13)[5]	Rs4973768	C/T (0.47)	1.16 (1.10–1.24)	5.8×10^{-7}
COX11	rs1156287	A/G (0.29)	0.91 (0.85–0.97)	0.0058	rs6504950	G/A (0.27)	0.91	0.95 (0.92–0.97)[5]	rs7222197 ($r^2 = 1.0$)	G/A (0.28)	0.92 (0.86–0.99)	0.021
RAD51L1	rs8009944	C/A (0.75)	0.88 (0.82–0.95)	0.0004	rs999737	C/T (0.24)	0.13	0.94 (0.88–0.99)[6]	rs999737	C/T (0.25)	0.89 (0.83–0.95)	0.0009
1p11.2	rs11249433	A/G (0.42)	1.08 (1.02–1.15)	0.010	rs11249433	A/G (0.39)	1.0	1.16 (1.09–1.24)[6]	rs11249433	A/G (0.42)	1.08 (1.02–1.15)	0.010
CASP8	rs10931936	T/C (0.74)	0.88 (0.82–0.94)	0.00015	rs1045485	G/C (0.13)	0.083	0.88 (0.84–0.92)[8]	rs17468277 ($r^2 = 1.0$)	C/T (0.13)	0.93 (0.85–1.02)	0.14

[a]Allele (frequency of the second listed allele). [b]Per-allele OR for the second listed allele, relative to the first. In each case the second listed allele was that which correlated with the second-listed published allele. [c]r^2 between the published SNP and most significant SNP in this study based on HapMap CEU. [d]r^2 between the published SNP and the best tagSNP in this study based on HapMap CEU. [e]Note that fine-mapping and functional analyses suggest that the strongest association for breast cancer is with rs2981578[25]. It is correlated with rs2981579 and rs2981582 at $r^2 = 0.85$. No more strongly correlated tag for rs2981578 was typed in the GWAS. [f]Estimated OR in Europeans. Estimated OR in Chinese was 1.36.

Table 2 Associations between genotype and breast cancer risk for six SNPs

Marker	Chromosome position	Stage[a]	Cases/controls	MAF[b]	Per-allele OR (95% CI)	Heterozygous OR (95% CI)	Homozygous OR (95%CI)	*P* value[c] Stage	*P* value[c] Combined
rs1011970 G/T	9 22,052,134	Stage 1	3,730/4,894	0.16	1.20 (1.11–1.30)	1.19 (1.08–1.31)	1.45 (1.13–1.86)	2.6×10^{-5}	
		Stage 2	12,253/12,000	0.17	1.09 (1.04–1.14)	1.07 (1.01–1.13)	1.29 (1.12–1.50)	0.00026	2.5×10^{-8}
rs2380205 C/T	10 5,926,740	Stage 1	3,730/4,895	0.44	0.86 (0.81–0.92)	0.86 (0.78–0.95)	0.75 (0.66–0.85)	7.9×10^{-5}	
		Stage 2	12,235/11,961	0.43	0.94 (0.91–0.98)	0.95 (0.90–1.01)	0.89 (0.82–0.95)	0.0017	4.6×10^{-7}
rs10995190 G/A	10 63,948,688	Stage 1	3,731/4,891	0.14	0.76 (0.70–0.84)	0.77 (0.69–0.86)	0.55 (0.40–0.77)	6.1×10^{-8}	
		Stage 2	12,261/12,000	0.15	0.86 (0.82–0.91)	0.84 (0.79–0.89)	0.83 (0.69–1.00)	1.4×10^{-8}	5.1×10^{-15}
rs704010 G/A	10 80,511,154	Stage 1	3,726/4,893	0.39	1.15 (1.09–1.23)	1.05 (0.95–1.15)	1.38 (1.22–1.57)	3.5×10^{-6}	
		Stage 2	12,222/11,992	0.39	1.07 (1.03–1.11)	1.11 (1.05–1.17)	1.13 (1.04–1.21)	0.00026	3.7×10^{-9}
rs614367 C/T	11 69,037,945	Stage 1	3,723/4,882	0.15	1.30 (1.20–1.41)	1.24 (1.13–1.37)	2.02 (1.56–2.64)	3.9×10^{-8}	
		Stage 2	12,114/11,967	0.15	1.15 (1.10–1.20)	1.16 (1.10–1.23)	1.27 (1.10–1.47)	1.3×10^{-8}	3.2×10^{-15}

[a]Stage 2 includes genotype data in SEARCH, RBCS and FBCS together with publicly available data from CGEMS. [b]MAF, frequency of the minor (second listed) allele. [c]Adjusted 1-d.f. *P* trend (see Online Methods).

a second stage involving 11,431 cases and 11,081 controls from four studies in the UK and The Netherlands (Online Methods). We also incorporated available data from 1,145 cases and 1,142 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) study. Six SNPs from five regions on chromosomes 9,10 and 11 showed clear evidence of replication in stage 2 ($P = 0.0017$ or better and in the same direction as stage 1) and reached significance levels over both stages combined of $P = 4.6 \times 10^{-7}$ to $P = 3.2 \times 10^{-15}$ (**Table 2** and **Supplementary Tables 2** and **3**). rs614367 and rs624797, which both showed strong evidence of association, were correlated, and rs624797 showed no independent association after adjustment for rs614367. The per-allele OR was higher in stage 1 than stage 2 for each SNP ($P < 0.05$ in each case; **Supplementary Table 2**). This may reflect either winner's curse or the enrichment of stage 1 for cases with a positive family history. There was no evidence for heterogeneity in the per-allele ORs among the stage 2 samples, with the exception of the weak evidence shown for rs10995190 ($P = 0.08$; **Supplementary Table 2**). There was no evidence for departure from a log-additive model for any SNP (that is, the OR for rare homozygotes did not differ significantly from the square of the OR for heterozygotes). There was weak evidence of a decrease in the per-allele OR with age for rs1011970 and of an increase in the per-allele OR with age for rs614367 ($P = 0.071$ and $P = 0.068$; **Supplementary Table 4**). rs614367 and rs624797 (but no other SNPs) showed a consistently stronger association with a positive family history in both stages (for rs614367, $P = 0.006$ and $P = 0.00016$, respectively; for rs624797, $P = 0.012$ and $P = 0.001$, respectively; **Supplementary Table 4**). For four of the SNPs (rs10995190, rs1011970, rs614367 and rs624797), the estimated per-allele ORs were higher for estrogen receptor–positive disease and showed little association in estrogen receptor–negative breast cancer, consistent with the pattern seen for the majority of breast cancer loci identified to date. For rs2380205 and rs704010, the per-allele ORs for estrogen receptor–positive and estrogen receptor–negative disease were similar, but the number of estrogen receptor–negative cases used was too small to draw firm conclusions on the effect sizes for this subset (**Supplementary Table 4**).

To examine whether there was evidence for a more strongly associated variant in any of the above regions, we used imputation to estimate the genotype probabilities in the stage 1 data at known SNPs in region using the HapMap CEU data as a framework. On chromosome 11, we identified four SNPs that showed a more significant association than rs614367 (most significantly associated SNP rs6610204; $P = 4.6 \times 10^{-14}$; **Supplementary Table 5**). In the other regions, no SNPs showed associations that were more significant than the original SNP. We also estimated the ORs associated with haplotypes of SNPs in each of the five regions (**Supplementary Table 6**). In each case, the association was present on more than one haplotype carrying the risk allele for the initially associated SNP, suggesting that the associations are unlikely to be driven by a single rare, high penetrance variant. For the chromosome 11 region, there was evidence of association with risk for two related haplotypes carrying the T allele of rs614367 with a combined frequency of 4%, suggesting that the causal variant may be somewhat rarer than the 15% minor allele frequency of rs614367.

SNP rs1011970 lies in a 180-kb block on 9p21 that includes *CDKN2A* and *CDKN2B*. These two genes encode cyclin-dependent kinase inhibitors and are frequently mutated or deleted in a wide variety of human tumors[11]. Germline mutations in *CDKN2A* predispose to malignant melanoma and pancreatic cancer[12], and recent GWAS also identified rs1011970 to be associated with melanoma risk[13]; SNPs within this same region are associated with nevus density and melanoma[14], basal cell carcinoma[15], glioma[16,17], diabetes[18] and coronary heart disease[19]. This is the first example of the same common variant predisposing to breast cancer and another cancer type rs10757278, which is correlated with rs1011970 ($r^2 = 0.7$), is associated with levels of expression in lymphocytes of *CDKN2A*, *CDKN2B* and a noncoding RNA in the same block, *CDKN2BAS* (also known as *ANR1L*)[20].

rs614367 on 11q13 lies in an LD block of ~166 kb that contains no annotated genes. This region is frequently amplified in human tumors, including breast cancers[21]. Plausible genes flanking this block include: proximally, *MYEOV*, a gene overexpressed in myeloma; distally, *CCND1*, encoding cyclin D1, a protein critical for cell-cycle control that is somatically altered in many tumor types; *ORAOV1*, a gene overexpressed in oral cancer; and three genes encoding fibroblast growth factors, *FGF19*, *FGF4* and *FGF3*. FGF3 and FGF4 are

oncogenic growth factors that bind distinct FGFR2 isoforms, providing a possible link with the *FGFR2* susceptibility locus[22].

rs10995190 on chromosome 10 lies within intron 4 of *ZNF365*, which encodes zinc finger protein 365. An amino acid substitution in this gene has been associated with uric acid nephrolithiasis[23]. Recent GWAS have identified another variant within this gene, rs10995271, located 159 kb downstream of rs10995190, to be associated with Crohn's disease[24]. rs2380205 lies in a 105-kb block on chromosome 10 containing the genes *ANKRD16* (encoding ankyrin repeat domain 16) and *FBXO18* (encoding the F-box protein, helicase 18). rs704010 on chromosome 10 lies in a 20-kb block 90 kb upstream of *ZMIZ1* (encoding zinc finger MIZ-type containing 1).

Based on the estimated per-allele ORs from stage 2 of our study, the newly identified loci explain approximately 1.2% of the familial risk of breast cancer, though the overall contribution may be larger because the true causal variants may be more strongly associated with disease than the SNPs tagging them in this study. Taken together with estimates from previous studies, the 18 confirmed breast cancer susceptibility loci explain approximately 8% of the familial risk of breast cancer, whereas rarer mutations in the known high risk loci (principally *BRCA1* and *BRCA2*) and moderate risk loci explain a further ~20%. This is by far the largest breast cancer GWAS to date and confirms that the *FGFR2* and *TOX3* loci (conferring per-allele ORs between 1.2 and 1.3) have the largest effect sizes from among the common susceptibility loci that are detectable with the current high-coverage genome-wide SNP sets. The residual familial risk is therefore likely to be due to a combination of a large number of common variants with smaller effects together with rarer variants not testable with current arrays. It is likely that many additional loci will be identifiable through more extensive follow-up of data from this and other GWAS.

URLs. CGEMS, http://cgems.cancer.gov/; Welcome Trust Case Control Consortium (WTCCC), http://www.wtccc.org.uk/; Nurses Health Study, http://www.channing.harvard.edu/nhs/; Mach, http://www.sph.umich.edu/csg/abecasis/MaCH/index.html; data access from this GWAS, http://www.srl.cam.ac.uk/genepi/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust and by Cancer Research UK. D.F.E. is a Principal Research Fellow of Cancer Research UK. C.T. is funded by a Medical Research Council Clinical Research Fellowship. The samples were collected and screened for *BRCA* mutations through funding from Cancer Research UK; US Military Acquis tion (ACQ) Activity, Era of Hope Award (W81XWH-05-1-0204) and the Institute of Cancer Research (UK). This study makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC) 2. A full list of the investigators who contributed to the generation of the data is available from the WTCCC website. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. Funding for this project was provided by the Wellcome Trust under award 085475. We thank the SEARCH team and Eastern Cancer Registry and Information Centre (ECRIC) for recru tment of the SEARCH cases. We acknowledge the clinicians from the Rotterdam Family Cancer Clinic who were involved in collecting the RBCS samples: C. Seynaeve, J. Klijn, J. Collee and R. Oldenburg.

AUTHOR CONTRIBUTIONS

D.F.E., N.R., M.R.S., P.D.P.P. and A.M.D. obtained funding for the study. D.F.E. designed the study and drafted the manuscript. D.F.E. and C.T. conducted the statistical analyses. J.M. provided data management and bioinformatics support. C.T. and N.R. coordinated the Familial Breast Cancer Study (FBCS). D.P., A.R., S.S., S.H., D.H., M.W.-P., C.T. and N.R. coordinated the FBCS genotyping. P.D.P.P. and D.F.E. coordinated Studies of Epidemiology and Risk Factors in Cancer Hered ty (SEARCH). S.A., M.M., M.G., C.S.H. and A.M.D. coordinated the stage 2 genotyping of the SEARCH and RBCS samples. M.H., M.S. and A.v.d.O. coordinated and provided samples and data from RBCS. C.L. coordinated the stage 1 genotyping. G.R. and R.H. provided data and samples from the Royal Marsden Hospital (RMH) study. W.T. and D.E. provided data and samples from the Prospective study of Outcomes in Sporadic vs. Hereditary breast cancer (POSH) study. D.E. and D.G.E. provided data and samples for FBCS. All authors provided critical review of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
2. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
3. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
4. Stacey, S.N. *et al.* Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **40**, 703–706 (2008).
5. Ahmed, S. *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* **41**, 585–590 (2009).
6. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41**, 579–584 (2009).
7. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**, 324–328 (2009).
8. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.* **39**, 352–358 (2007).
9. Antoniou, A.C. & Easton, D.F. Polygenic inheritance of breast cancer: implications for design of association studies. *Genet. Epidemiol.* **25**, 190–202 (2003).
10. Milne, R.L. *et al.* Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. *J. Natl. Cancer Inst.* **101**, 1012–1018 (2009).
11. Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* **264**, 436–440 (1994).
12. Kamb, A. *et al.* Analysis of the p16 gene (*CDKN2*) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat. Genet.* **8**, 23–26 (1994).
13. Bishop, D.T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat. Genet.* **41**, 920–925 (2009).
14. Falchi, M. *et al.* Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. *Nat. Genet.* **41**, 915–919 (2009).
15. Stacey, S.N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat. Genet.* **41**, 909–914 (2009).
16. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899–904 (2009).
17. Wrensch, M. *et al.* Variants in the *CDKN2B* and *RTEL1* regions are associated with high-grade glioma susceptibility. *Nat. Genet.* **41**, 905–908 (2009).
18. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
19. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
20. Liu, Y. *et al.* INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. *PLoS One* **4**, e5027 (2009).
21. Karlseder, J. *et al.* Patterns of DNA amplification at band q13 of chromosome 11 in human breast cancer. *Genes Chromosom. Cancer* **9**, 42–48 (1994).
22. Ornitz, D.M. *et al.* Receptor specificity of the fibroblast growth factor family. *J. Biol. Chem.* **271**, 15292–15297 (1996).
23. Gianfrancesco, F. *et al.* Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. *Am. J. Hum. Genet.* **72**, 1479–1491 (2003).
24. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
25. Udler, M.S. *et al.* FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum. Mol. Genet.* **18**, 1692–1703 (2009).

ONLINE METHODS

Samples. Three thousand nine hundred and sixty breast cancer cases were used in stage 1, of which 3,652 were from cancer genetics clinics in the UK recruited through the Familial Breast Cancer Study (FBCS) and 308 were from oncology clinics in the UK recruited through the Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) study. Cases were preferentially selected to have at least two affected first- or second-degree relatives. The majority of cases were screened and found to be negative for germline mutations, including large rearrangements, in *BRCA1* and *BRCA2*. A minority of samples were not tested for *BRCA1* or *BRCA2* mutations. All carriers of disease-associated mutations in *BRCA1* and *BRCA2* were excluded. We also excluded all cases with self-reported non-European ancestry.

Controls for stage 1 were drawn from two sources: 2,930 controls were drawn from the 1958 Birth Cohort (1958BC), a population-based study in the United Kingdom of individuals born in 1 week in 1958 (ref. 26). The remaining 2,737 controls were identified through the UK National Blood Service (NBS)[19]. These samples were genotyped as part of the Wellcome Trust Case Control Consortium (WTCCC2; see URLs)[27]. The analyses presented here are based on 2,482 1958BC and 2,587 NBS controls for which genotype data were available at the time of analysis.

Samples for stage 2 were drawn from six sources: (i) the SEARCH study, a population-based study of cases from East Anglia (n = 6,640); controls (n = 6,832) were drawn from the European Prospective Investigation into Cancer and Nutrition (EPIC) study, a population-based cohort study of diet and cancer from general practices contributing to SEARCH; (ii) the Rotterdam breast cancer study (RBCS) (799 cases, 800 controls); (iii) the Familial Breast Cancer Study (FBCS), consisting of additional cases ascertained through UK cancer genetics clinics (n = 2,009); (iv) the RMH breast cancer series (n = 1,732); and (v) the Prospective study of Outcomes in Sporadic vs. Hereditary breast cancer (POSH) study (n = 251). The combined samples from these latter three series (n = 3,992) were analyzed in a single replication experiment together with additional controls selected through the 1958BC (n = 3,450), none of which were included in stage 1. For stage 2, we also incorporated data on the relevant SNPs from the CGEMS study (see URLs). CGEMS is based on 1,145 cases and 1,142 controls drawn from the Nurses Health Study (see URLs) which were genotyped using the Illumina 550k array.

All studies were approved by the appropriate ethics committees.

Genotyping. Genotypes for stage 1 cases were generated using a custom Illumina Infinium 670k array and controls were genotyped using an Illumina Infinium 1.2M array at the Wellcome Trust Sanger Institute. For this analysis, we analyzed data on 594,375 SNPs that were successfully genotyped on both arrays. Genotypes for both arrays were called using the Illuminus algorithm[28]. We used genotypes for which Illuminus generated a posterior probability of >0.95. Cluster plots were inspected manually for all SNPs considered for inclusion in stage 2.

Genotyping for stage 2 was performed by 5′ exonuclease assay (Taqman) using the ABI Prism 7900HT Sequence Detection System according to the manufacturer's instructions. Primers and probes were supplied directly by Applied Biosystems as Assays-By-Design. Assays included at least two negative controls and 2% to 5% duplicates per plate. Genotyping for one marker, rs1866823, failed for the SEARCH and RBCS studies, and the marker was replaced by rs2246873 (r^2 = 0.94 in HapMap CEU).

Analyses. We restricted analyses to individuals who were called on >97% of successfully genotyped SNPs. To identify close relatives, we computed identity-by-state (IBS) probabilities for all pairs. We confirmed 2 case monozygotic twin (MZ) pairs, 22 duplicate case pairs and 24 first-degree relative pairs (IBS > 0.86). We also identified 4 probable case-control and 44 probable control-control sibling pairs. We excluded the control from the case-control pairs and the sample with the lower call rate from the remaining pairs. By computing IBS scores between participants and individuals in HapMap and by using multidimensional scaling, we identified 89 individuals who appeared to have substantial Asian or African ancestry (defined as approximately >15% non-European ancestry, comprising 69 cases, 4 individuals from 1958BC and 16 NBS). After these exclusions, 3,659 cases and 4,897 controls were used in the final analysis.

We filtered out all SNPs with, in either cases or controls, a MAF < 1%, a call rate of < 99% and a MAF < 5%, or a call rate < 95% and MAF ≥ 5%. We also excluded SNPs whose frequencies departed from HWE at $P < 0.00001$ in controls or $P < 10^{-12}$ in cases. After these exclusions, we used data on 582,886 SNPs. Duplicate concordance was 99.99%.

Statistical methods. We first assessed associations between each SNP and breast cancer at stage 1 using a 1-d.f. Cochran-Armitage trend test and a general 2-d.f. χ^2 test. Inflation in the χ^2 statistic was assessed using the genomic control approach; we derived an inflation factor λ by dividing the median of the lowest 90% of the 1-d.f. statistics by the 45% percentile of a 1-d.f. χ^2 distribution (0.357). We have also presented the equivalent inflation factor for a study of 1,000 cases and 1,000 controls ($\lambda_{1,000}$) calculated by $\lambda_{1,000} = 1 + 500\,(1 / N_{cases} + 1 / N_{controls}) / (\lambda - 1)$, where N_{cases} and $N_{controls}$ are the number of cases and controls, respectively.

To correct for potential inflation due to population structure, we performed a principal-components analysis based on the genotypes of a subset of 35,797 uncorrelated SNPs ($r^2 < 0.1$)[29]. We then computed 1-d.f. score tests for each SNP, adjusting for progressively larger numbers of principal components as covariates. Adjustment for the first two components reduced the inflation slightly (to 1.06); however, adjustment for further components did not reduce the inflation further. Adjusted significance tests were therefore calculated from the score tests adjusted for two principal components. To allow for the residual inflation, we adjusted the resulting test statistics using the genomic control approach by dividing the test statistic by the inflation factor.

SNPs were selected for evaluation in stage 2 on the basis of a significance level of $P < 10^{-5}$ based on the unadjusted 1-d.f. trend test. Where two or more SNPs were selected from the same region, we used multiple logistic regression to determine a minimal set of SNPs that showed evidence of association after adjustment for other SNPs. In practice, one SNP was selected in each region except in the case of one region, in which two SNPs were genotyped.

After stage 2, overall 1-d.f. and 2-d.f. tests of association were derived, stratified by stage and study. Adjusted tests of association were derived by adjusting in stage 1 for principal components and genomic control as described above. In the combined analysis, the effect size in stage 1 was weighted by a factor of 2 relative to that in stage 2, consistent with the effect size expected under a polygenic model. A criterion of $P < 5 \times 10^{-7}$ was used for genome-wide significance[19], and ORs and 95% confidence limits were estimated using unconditional logistic regression, stratified by study. In the text, we have reported the combined tests of association over both stages, but we have emphasized the OR estimates from stage 2 to minimize the effect of winner's curse. Tests of homogeneity of the ORs across strata were assessed using likelihood ratio tests. The associations between genotype and family history in stage 2, and between genotype and estrogen receptor status, were assessed using a case-only analysis (that is, treating family history or estrogen receptor status as the outcome variable and estimating a per-allele OR for each SNP using logistic regression). For stage 1, the effect of family history was analyzed using a family history score, derived as the total number of affected relatives weighted by their degree of relationship to the case. The effect of family history score on per-allele OR was assessed using constrained polytomous regression. Modification of the ORs by age at diagnosis was assessed using a case-only analysis, assessing the association between age and SNP genotype in the cases using polytomous regression. The contribution of the loci to the familial risk of breast cancer was estimated by first computing the familial risk to a daughter of an affected individual that was attributable to each locus (λ_1) from the allele frequency and the estimated per-allele OR in the SEARCH study, which was largest study contributing to stage 2 and which is population based. The proportion of the familial risk due to each locus was then calculated as $\ln(\lambda_1) / \ln(2)$, assuming an overall familial relative risk of 2. The combined effect of all loci was then derived by summing the locus-specific contributions (that is, assuming a log-additive model). Imputed genotypes for non-typed SNPs were estimated using Mach (see URLs), using the HapMap CEU data as a framework. Haplotype analyses were conducted in haplo.stats[30]. Haplotypes were based on SNPs in each region that were significantly associated with breast cancer at $P < 0.001$, after eliminating perfectly correlated SNPs. Per-haplotype ORs were estimated using the haplo.cc routine. Other analyses were performed in R, principally using GenABEL[31], and Stata (R, http://www.r-project.org/; Stata, http://www.stata.com/).

26. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35**, 34–41 (2006).
27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
28. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746 (2007).
29. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
30. Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. & Poland, G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).
31. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

doi:10.1038/ng.586

PERSPECTIVE

nature genetics

The emerging landscape of breast cancer susceptibility

Michael R Stratton & Nazneen Rahman

The genetic basis of inherited predisposition to breast cancer has been assiduously investigated for the past two decades and has been the subject of several recent discoveries. Three reasonably well-defined classes of breast cancer susceptibility alleles with different levels of risk and prevalence in the population have become apparent: rare high-penetrance alleles, rare moderate-penetrance alleles and common low-penetrance alleles. The contribution of each component to breast cancer predisposition is still to be fully explored, as are the phenotypic characteristics of the cancers associated with them, the ways in which they interact, much of their biology and their clinical utility. These recent advances herald a new chapter in the exploration of susceptibility to breast cancer and are likely to provide insights relevant to other common, heterogeneous diseases.

In most Western populations, approximately one in ten women develop breast cancer. Epidemiological studies have shown that first-degree female relatives of women with breast cancer are at approximately two-fold risk of developing the disease compared to the general population[1]. Although, in principle, this could be attributable to shared environmental or genetic factors, or both, twin studies indicate that most of the excess familial risk is due to inherited predisposition[2].

Rare high-penetrance breast cancer susceptibility genes

Major advances in understanding breast cancer susceptibility were made in the last decade of the twentieth century through genetic linkage mapping and positional cloning of two major predisposition genes, *BRCA1* and *BRCA2* (refs. 3–6). Disease-causing variants in *BRCA1* and *BRCA2* confer a high risk of breast cancer, approximately 10- to 20-fold relative risk. This translates into a 30–60% risk by age 60, compared to 3% in the general population. The relative risks are higher for early-onset breast cancers, and there are also elevated risks of ovarian and other cancers[7,8]. Disease-causing mutations in *BRCA1* and *BRCA2* result in inactivation of the encoded proteins, generally by causing premature protein truncation or nonsense-mediated RNA decay. There is population variation in mutation prevalence, but mutations are infrequent in most populations. Approximately 1 in 1,000 individuals in the UK are heterozygous mutation carriers of each gene, and there are numerous different mutations, each of which is very rare[9,10]. Cancer predisposition is transmitted as an autosomal dominant trait in families harboring mutations. However, at the cellular level, *BRCA1* and *BRCA2* act as recessive cancer genes, with mutations converted to homozygosity in the cancers which they cause, usually through loss of the wild-type allele. Several years of biological investigation have firmly implicated BRCA1 and BRCA2 in double-strand DNA break repair[11].

Mutations in *BRCA1* and *BRCA2* account for ~16% of the familial risk of breast cancer[9,10]. Germline mutations in *TP53* cause Li-Fraumeni syndrome, which includes a high risk of breast and other cancers, but these mutations are very rare and hence account for a much smaller proportion of the familial risk. Cancer predisposition syndromes due to mutations in *PTEN* (Cowden syndrome), *STK11* (Peutz-Jeghers syndrome) and *CDH1* are also associated with elevated risks of breast cancer, although the cancer risks and prevalence of mutations in these genes are not well defined. It is unlikely that mutations in all six of these genes together account for more than 20% of the familial risk of the disease[12,13]. Genome-wide linkage analyses using large numbers of families without mutations in *BRCA1* or *BRCA2* have not mapped additional susceptibility loci[14]. Although this does not completely exclude the existence of further high-penetrance breast cancer susceptibility genes, it strongly suggests that, if they exist, they account for a very small fraction of familial risk. So, how can the remaining ~80% of the familial risk of breast cancer be explained?

A new harvest of breast cancer susceptibility alleles has recently emerged through two distinct strategies: direct interrogation of genes believed to be strong candidates, which has led to the identification of rare moderate-penetrance alleles[15–19], and genome-wide tag SNP association studies, which have identified common low-penetrance alleles[20–22] (**Box 1**). We have considered these two new classes separately and in distinction to the rare high-penetrance genes discussed previously. It is possible that the differences among these classes may, at least in part, be attributable to the methods employed in their identification, and further discoveries may render the boundaries among them less distinct. Nevertheless, they currently provide a useful basis for considering the genetic landscape of breast cancer susceptibility.

Rare moderate-penetrance breast cancer susceptibility genes

The candidacy of the breast cancer susceptibility genes recently identified through direct interrogation for disease-causing mutations has been

Michael R. Stratton and Nazneen Rahman are at the Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. Michael R. Stratton is in the Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. e-mail: nazneen.rahman@icr.ac.uk

Published online 27 December 2007; doi:10.1038/ng.2007.53

Box 1 Classes and key features of known breast cancer susceptibility alleles

High-penetrance breast cancer susceptibility genes

Examples: *BRCA1*, *BRCA2*, *TP53*

- **Risk variants:** Multiple, different mutations that predominantly cause protein truncation
- **Frequency:** Rare (population carrier frequency ≤0.1%)
- **Risk of breast cancer:** 10- to 20-fold relative risk
- **Primary strategy for identification:** Genome-wide linkage and positional cloning

Moderate-penetrance breast cancer susceptibility genes

Examples: *ATM*, *BRIP1*, *CHEK2*, *PALB2*

- **Risk variants:** Multiple, different mutations that predominantly cause protein truncation
- **Frequency:** Rare (population carrier frequency ≤0.6%)
- **Risk of breast cancer:** two- to fourfold relative risk
- **Primary strategy for identification:** Direct interrogation of candidate genes for coding variants in large, genetically enriched breast cancer case series and controls

Low-penetrance breast cancer susceptibility alleles

Examples: rs2981582 (*FGFR2*, 10q), rs3803662 (*TNRC9* (recently renamed *TOX3*), 16q), rs889312 (*MAP3K1*, 5q), rs3817198 (*LSP1*, 11p), rs13281615 (8q), rs13387042 (2q), rs1045485 (*CASP8*_D302H)

- **Risk variants:** Single-nucleotide polymorphisms that are causal or in linkage disequilibrium with the causal variant(s). May occur in noncoding, nongenic regions.
- **Frequency:** Common (population frequency 5–50%)
- **Risk of breast cancer**: up to ~1.25-fold (heterozygous) or 1.65-fold (homozygous) relative risk
- **Primary strategy for identification:** Genome-wide association studies of hundreds of thousands of SNPs in large breast cancer case-control series

based primarily on involvement of the encoded proteins in biological pathways that include BRCA1 and BRCA2. To date, this strategy has identified at least four genes: *CHEK2*, *ATM*, *BRIP1* and *PALB2* (refs. 15–19). CHEK2 is a checkpoint kinase involved in DNA repair that directly modulates the activities of p53 and BRCA1 by phosphorylation[23]. *ATM* also encodes a checkpoint kinase that has key functions in DNA repair, and which also phosphorylates p53 and BRCA1 (ref. 24). BRIP1 (also known as BACH1) was discovered as a binding partner of BRCA1 and is implicated in some BRCA1 activities relating to DNA repair[25]. PALB2 was discovered as a protein associated with BRCA2 (ref. 26). The patterns of susceptibility associated with these four genes have many features in common.

In *CHEK2*, *ATM*, *BRIP1* and *PALB2*, most of the disease-causing mutations result in premature protein truncation or nonsense-mediated RNA decay through nonsense codons or translational frameshifts. A small proportion is likely to be rare missense variants that disrupt critical functions. In each of the four genes, there are multiple different pathogenic mutations, each of which is generally very rare. Disease-causing mutations in each gene are found in less than 1% of the UK population: ~0.6% are heterozygous carriers of *CHEK2* mutations (a single mutation, *CHEK2**1100delC, accounts for most of these), ~0.4% are heterozygous carriers of *ATM* mutations and ~0.1% or fewer are heterozygous carriers of *BRIP1* or *PALB2* mutations[15–18,27]. The prevalence of mutations in most other populations is currently less well characterized, although it is noteworthy that founder mutations in *CHEK2* and *PALB2* in Finland allowed independent identification of the association of these genes with breast cancer[19,28].

Overall, with respect to their effect on protein function, their prevalence in the population and their biological consequences, disease-causing mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* bear many similarities to disease-causing mutations in *BRCA1* and *BRCA2*. Where they differ is in the risks of breast cancer they confer. Although there is currently some imprecision in the risk estimates, it is clear that mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* confer less elevated risks of breast cancer (about two- to threefold, with confidence intervals ranging from 1.2 to 3.9) than mutations in *BRCA1* or *BRCA2* (10- to 20-fold)[15–18,27]. Carriers of moderate-penetrance mutant alleles therefore have approximately a 6–10% risk of developing breast cancer by age 60, compared to ~3% in the general population. For each gene, it is possible that there is risk heterogeneity, with some variants conferring greater risks than others (as is the case for *BRCA1* and *BRCA2* mutations), but there are currently few persuasive examples of this. Because *CHEK2*, *ATM*, *BRIP1* and *PALB2* mutations confer a smaller increased risk of breast cancer than *BRCA1* and *BRCA2* mutations, and their disease-causing mutations are uncommon, each of these moderate-risk genes makes a relatively small contribution to the overall familial risk of breast cancer. Current estimates suggest that mutations in the four genes together account for 2.3% of the familial risk of breast cancer, compared to 16% for *BRCA1* and *BRCA2* together[9,10,12,15].

Features of rare moderate-penetrance susceptibility genes

Despite the many similarities of *CHEK2*, *ATM*, *BRIP1* and *PALB2* to *BRCA1* and *BRCA2*, the lower breast cancer risk conferred by mutations in the former group leads to some uncomfortable departures from familiar genetic patterns. For example, in breast cancer–affected families carrying *BRCA1* or *BRCA2* mutations, the mutation and disease status usually track together, although even in this context the occasional sporadic 'phenocopy' is encountered. However, when the breast cancer risks associated with a particular allele are only two- to threefold, disease-causing mutations often do not segregate with the disease. This is because most mutation carriers do not actually develop breast cancer, because the sporadic rate of breast cancer is high, and because familial breast cancer clusters not associated with mutations in *BRCA1* or *BRCA2* probably reflect chance aggregations of susceptibility alleles in multiple different genes. As a consequence, segregation of the disease with the mutation, which is one of the tests a new disease susceptibility gene is routinely subjected to, is generally unhelpful for confirmation of lower-penetrance alleles. If sufficient multiply sampled breast cancer–affected families with mutations are analyzed, it should be possible to formally show that the mutation segregates with the disease more frequently than

would occur simply by chance. Thus far, however, sufficient families have only been available to show this for *CHEK2* (ref. 16).

Similarly, the familiar pattern of loss of the wild-type allele in cancers, which is generally associated with high-penetrance autosomal dominant cancer genes that operate in a recessive fashion in cancer cells, may be less apparent when sought in the context of lower-penetrance susceptibility alleles. Given the predominant pattern of inactivating disease-causing mutations, it is mechanistically plausible that *CHEK2*, *ATM*, *BRIP1* and *PALB2* behave in a fashion similar to *BRCA1* and *BRCA2* and show somatic loss of the wild-type allele in the cancers they cause. However, to demonstrate this pattern may require analysis of a substantial number of tumors, because only about half of breast cancers in individuals with a mutation in a cancer susceptibility gene conferring a twofold risk arise because of the mutation—the remainder would have occurred anyway. Allelic loss in cancers not due to the mutation will follow the pattern present in sporadic cancers for that locus, and will target the wild-type and mutant alleles equally. Thus, it may be necessary to analyze a large series of breast cancers from mutation carriers before meaningful, statistically robust data on loss of the wild-type allele can be obtained.

Elucidation of the phenotypes associated with heterozygous mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* will also be hindered by the considerations discussed above, compounded by the rarity of disease-causing alleles. At this stage, strong evidence does not exist for a higher risk of early-onset breast cancer, but most studies have had insufficient power to demonstrate it. The risks of other cancers, and the histological phenotypes of the breast cancers associated with mutations in these genes, are uncertain and may require large-scale collaborative initiatives to generate sufficient numbers.

Phenotypes associated with biallelic mutations

Mutations in high- and moderate-penetrance breast cancer genes confer an elevated risk of breast cancer in monoallelic (heterozygous) carriers. However, individuals with biallelic (homozygous or compound heterozygous) mutations in some of these genes have a different phenotype, often manifesting during childhood. This is exemplified by *ATM*, which was initially discovered by positional cloning of the gene underlying ataxia telangiectasia, an autosomal recessive condition characterized by loss of cerebellar Purkinje cells, immune deficiency and cancer predisposition[29]. Several epidemiological studies over the past two decades have shown that heterozygous (monoallelic) female carriers of ataxia telangiectasia–causing *ATM* mutations are at elevated risk of breast cancer, and molecular confirmation of this association was finally reported last year[17,30].

Similarly, in 2002, it was shown that biallelic *BRCA2* mutations cause a rare subgroup of Fanconi anemia, subtype FA-D1 (ref. 31). Fanconi anemia is a genetically heterogeneous, recessive, chromosomal instability disorder characterized by growth retardation, skeletal abnormalities, bone marrow failure, cancer predisposition and cellular hypersensitivity to DNA cross-linking agents. FA-D1 is a distinctive subtype associated with severe disease and a high risk of childhood solid tumors such as Wilms tumor, medulloblastoma and glioma that occur rarely in classic Fanconi anemia[32]. Subsequently, it was shown that biallelic mutations in *BRIP1* and *PALB2* also cause rare subgroups of Fanconi anemia (FA-J and FA-N, respectively)[33–36]. The phenotype of FA-N, resulting from biallelic *PALB2* mutations, is characterized by severe disease and a high risk of childhood solid tumors and is virtually identical to that of FA-D1, presumably reflecting the close functional relationship between BRCA2 and PALB2 (refs. 32,34). However, FA-J, caused by biallelic *BRIP1* mutations, results in the classic Fanconi anemia phenotype and has not been associated with childhood solid tumors[33,36]. It is possible that biallelic mutations in additional breast cancer susceptibility genes are responsible for other Fanconi anemia subtypes. However, both epidemiological and molecular analyses suggest that only a subset of Fanconi anemia genes are breast cancer susceptibility genes[37]. The factors that determine whether a Fanconi anemia gene is also a breast cancer predisposition gene are not known.

There is no known phenotype associated with biallelic mutations in *CHEK2* or *BRCA1*. One individual homozygous for *CHEK2**1100delC has been reported and was healthy until developing colorectal cancer at 52 years[38]. Conversely, although more than a decade has elapsed since *BRCA1* was identified, no confirmed *BRCA1* biallelic mutation carrier has been reported. It is conceivable that biallelic *BRCA1* mutations cause a rare syndrome yet to be attributed to this gene, are embryonic lethal or (perhaps less likely) are not associated with any distinctive phenotype.

Common low-penetrance breast cancer susceptibility alleles

A third component of the landscape of breast cancer susceptibility has been the subject of speculation for years, but has only just begun to surface. It is comprised of common alleles that confer very small increases in risk (common low-penetrance alleles). The currently known susceptibility alleles of this type have been discovered through association studies, either targeted at individual genes on the basis of biological candidacy or, more recently, through genome-wide tag SNP searches. In the past, numerous associations were proposed from targeted association studies involving relatively small numbers of cases and controls. Most of these have not been confirmed when evaluated on additional series, and such observations have acquired a certain notoriety and disrepute. Progress in this area of breast cancer research has depended, at least in part, on the formation of multigroup collaborations that combine data from very large numbers of cases and controls from many different locations and ethnic groups. These combined sets of tens of thousands of cases and controls provide substantial power to detect small effects and can obviate problems and limitations intrinsic to individual series[39].

Only a small number of statistically unimpeachable, common low-penetrance breast cancer susceptibility alleles have thus far been reported and confirmed in different populations[20–22]. For the purposes of this review, we focus on seven for which there is strong evidence and that can serve to illustrate at least the outlines of the emerging landscape[20–22,40]. However, these are unlikely to represent all the patterns that will be found in future studies.

Five of the seven confirmed breast cancer risk alleles are within regions of linkage disequilibrium that cover known protein-coding genes. The genes in these regions include *CASP8* (encoding caspase 8, a member of the cysteine-aspartic acid protease family whose sequential activation has a central role in the execution of apoptosis), *FGFR2* (encoding fibroblast growth factor receptor 2), *TNRC9* (recently renamed *TOX3*, encoding a protein with a putative high-mobility-group motif suggesting that it might act as a transcription factor), *MAP3K1* (encoding mitogen-activated protein kinase kinase kinase 1, a protein likely involved in growth signaling) and *LSP1* (encoding lymphocyte-specific protein 1, an intracellular F-actin binding protein). Some of these regions of linkage disequilibrium contain other genes, and it is conceivable that the functional associations are related to these rather than to the genes cited above, or perhaps to other, currently cryptic, genetic elements. Two of the seven susceptibility loci are on 8q and 2q, in regions with no known protein-coding genes[20–22,40].

The increased risks of breast cancer conferred by these seven susceptibility alleles are small. The relative risks of breast cancer associated with carrying a single copy of each risk allele range from 1.07 to 1.26, with the *FGFR2* and 2q susceptibility alleles at the high end of this spectrum. The population prevalence of each risk allele is high, however, ranging from 28% to 87%. Interestingly, for some of these loci, the higher-risk

allele is the more common. Because the predisposing alleles are common, despite the low risks they confer, their contribution to the familial risk of breast cancer is relatively substantial. The six loci characterized by Easton *et al.* and Cox *et al.* are estimated to account for 3.9% of the familial risk of breast cancer in European populations[20,40].

It is likely that there are very few, if any, additional common low-penetrance susceptibility alleles that make contributions to the familial risk of breast cancer as substantial as those in *FGFR2* or the locus on 2q. However, there is evidence for the existence of many, perhaps hundreds of, yet-to-be-discovered common susceptibility alleles with smaller effects[20]. Therefore, a sizeable proportion of the genetic architecture of breast cancer susceptibility may be embodied in a multitude of common susceptibility alleles, each of which accounts for a very small fraction of the familial risk.

Features of common low-penetrance susceptibility alleles

The disease-causing variants underlying these recently reported associations may not be easily identifiable, because the primary association is with a sentinel, reporter SNP that is often in tight linkage disequilibrium with many nearby variants. Even if the disease-causing variant is ultimately identified, it may not be obvious which gene(s) mediates its biological effects. Despite these complications and the limited number of common low-penetrance breast cancer susceptibility alleles thus far identified, some incipient trends and patterns may be emerging.

First, common low-penetrance breast cancer risk variants frequently reside in noncoding regions of the genome. For example, the susceptibility variant in *FGFR2* is within an intron of the gene. Moreover, the susceptibility variants on 2q and 8q are both several tens of kilobases away from the nearest protein-coding genes. Of particular interest is the locus on 8q, which is in close proximity to different linkage disequilibrium blocks that contain alleles predisposing to prostate cancer and colorectal cancer[41–47]. It seems unlikely that this physical clustering is simply coincidence. Nevertheless, it remains to be seen whether these associations are mediated by a related biological mechanism.

Second, the mechanism of action of at least some common low-risk breast cancer–predisposing loci may be through activation of growth-promoting genes, in contrast to the inactivation of DNA repair genes that characterizes known rare high- and moderate-risk genes. For example, somatically acquired missense mutations, amplification and overexpression of *FGFR2* are well documented in human cancer and result in overactivity of the protein[48,49]. Furthermore, the gene closest to the breast, prostate and colorectal cancer risk variants on 8q, remarkably, is *MYC*, which is commonly amplified or overexpressed through chromosomal rearrangement in many types of cancer. Assuming that the predisposing variants at these loci are exerting their effects through FGFR2 and MYC (which is by no means certain), our current understanding of these genes would predict that the susceptibility alleles increase the activity of the encoded proteins. However, most of the currently mapped common low-penetrance loci are anonymous or have functions previously unrelated to cancer development, and they therefore may lead us into previously uncharted areas of cancer biology.

Third, in contrast to the rare high-penetrance and moderate-penetrance genes, homozygosity for a common low-penetrance susceptibility variant does not usually confer a distinct phenotype. Instead, homozygotes are phenotypically normal, but have an increased breast cancer risk that seems to be approximately the product of the risk for heterozygotes. Exploration of the histological phenotypes of cancers associated with common low-penetrance alleles is in its infancy, although at least some of these alleles seem to be particularly associated with estrogen receptor–positive breast cancers, in contrast to *BRCA1* mutations, which are strongly associated with estrogen receptor–negative tumors[22,50].

Identification of further breast cancer susceptibility genes

The recent discoveries described here have together exposed a clearer picture of the genetic architecture of breast cancer susceptibility. *BRCA1* and *BRCA2* are likely to be the only major high-penetrance breast cancer susceptibility genes, and together with other rare, high-penetrance genes, they account for approximately 20% of the familial risk of disease. The remaining susceptibility is therefore due to genes conferring more modest increases in risk. *CHEK2*, *ATM*, *BRIP1* and *PALB2* are breast cancer susceptibility genes that bear many biological similarities to *BRCA1* and *BRCA2* but confer a breast cancer relative risk of two- to fourfold. They represent the current paradigms for a second class of rare moderate-penetrance risk alleles, but it would not be surprising if other such genes exist.

As disease-causing mutations in these genes do not generally result in large pedigrees with multiple breast cancer cases, further susceptibility genes of this class will not easily be mapped by genetic linkage analysis. Moreover, because the disease-causing alleles are uncommon, it is unlikely that they will be detected by association studies. Therefore, the most effective strategy to detect this class of gene is likely to remain the systematic screening of entire genes for potential disease-causing variants (usually truncating mutations) in series of breast cancer cases compared to controls. Because the breast cancer risks conferred by these variants are only two- to fourfold and the risk alleles are rare, the numbers of subjects required in these studies are large, rendering the analyses laborious by current technology. The problem can, to some extent, be mitigated by using familial rather than population-based breast cancer cases, as even lower-penetrance breast cancer susceptibility alleles are usually enriched in familial breast cancer cases compared to nonfamilial series. Use of population isolates with founder mutations of higher prevalence than is typical of outbred populations can also empower gene identification studies[19]. Such studies in Finnish breast cancer cases have provided suggestive data that *RAD50* may be a moderate-penetrance breast cancer predisposition gene, although the rarity of truncating mutations precluded confirmation of an association with breast cancer in UK families[51,52]. It is difficult to predict how many more rare moderate-penetrance genes exist, how much breast cancer susceptibility is accounted for by this component of the landscape or whether this pattern of susceptibility will extend beyond genes involved in DNA repair. Furthermore, the resequencing studies required for their identification are currently restricted to limited sets of candidate genes. However, with the likely advent of genome-wide resequencing of constitutional DNA, further exploration of this class of susceptibility allele should be possible.

Finally, the floodgates seem to be opening for the set of common low-penetrance alleles that confer risks of 1.3-fold or less. Although the current state of knowledge is sketchy, we can at least now be sure that they exist and that they show biological differences from the rare high-penetrance and rare moderate-penetrance genes. Only a small proportion of the familial risk of breast cancer is thus far explained by well-supported examples of this class of susceptibility allele. However, it is possible that a substantial proportion of the still unexplained (>70%) familial risk may be due to large numbers of similar variants with smaller effects. Further studies should yield additional variants in this class, although even with existing large-scale collaborations, sufficient samples may not yet be available to conclusively identify many variants with weak effects.

Are there other areas of the landscape to be explored? An intriguing feature is the apparent discontinuity of breast cancer risks among the three currently defined groups of susceptibility alleles. Mutations in *BRCA1* and *BRCA2* confer 10- to 20-fold relative risks of breast cancer, the rare moderate-penetrance genes confer relative risks of 2- to 4-fold and the common low-penetrance alleles confer relative risks less than

1.3-fold. Whether this pattern reflects a genuine biological stratification or an ascertainment artifact compounded by the limited number of known alleles remains to be seen.

It is also plausible that rare, nontruncating variants contribute to the genetic architecture of breast cancer susceptibility, given that rare truncating and common nontruncating variants are already known to be important. Investigating the role of rare nontruncating variants will, however, be challenging; their rarity will severely hamper detection through association studies, and it is very difficult to distinguish pathogenic nontruncating variants a priori from the plethora of innocuous rare variants.

Interactions between breast cancer susceptibility alleles

The available data suggest that many familial breast cancer clusters are likely to be due to the coincidence of multiple, lower-risk breast cancer susceptibility alleles[13,53]. This raises the question of the manner in which each breast cancer susceptibility allele in such clusters interacts with the others. The evidence for the common low-penetrance variants seems to indicate that, in general, they interact with each other multiplicatively[20,22]. Investigation of the breast cancer risks conferred by *CHEK2**1100delC, however, showed that the pattern of multiplicative interaction does not always apply. Although *CHEK2**1100delC confers an approximately twofold risk of breast cancer in most genetic backgrounds, it does not seem to confer an elevated breast cancer risk in carriers of *BRCA1* or *BRCA2* mutations[16]. Understanding that the proteins encoded by these genes lie in the same biological pathways provides a simple but credible explanation. In this example, abrogation of functions of these pathways by an inactivating mutation of *BRCA1*, *BRCA2* or *CHEK2* confers breast cancer susceptibility. However, if the relevant function is already abolished by a *BRCA1* or *BRCA2* mutation, an inactivating mutation in *CHEK2* will not confer an additional breast cancer risk. Because CHEK2 is known to phosphorylate and regulate BRCA1 and is involved elsewhere in double-strand DNA break repair, this notion has a reasonably solid foundation in our current understanding of these pathways[11,23].

It is currently unknown how common susceptibility alleles interact with rare susceptibility variants, though it is likely that relevant data will be forthcoming in the near future. Exploration of interactions among breast cancer risk alleles and nongenetic factors, such as hormonal profiles and environmental exposures, is also in its infancy, and will be vital in building a comprehensive picture of the underlying causes of familial clustering of the disease.

Clinical utility

Diagnostic testing for mutations in *BRCA1* and *BRCA2* has been routine clinical practice in many countries for several years. It facilitates risk estimation and implementation of cancer prevention strategies and increasingly has the potential to influence cancer therapy[54,55]. Management interventions in breast cancer–affected families without *BRCA1* or *BRCA2* mutations have inevitably been more limited, as less information has been available for risk evaluation. The identification of new susceptibility alleles may offer the potential for improved care in such families: for example, if combinations of alleles alter the risk category of an individual such that screening or prophylactic interventions might be considered. However, clinical testing of the new generation of susceptibility genes will need to be undertaken carefully and cautiously, and more detailed information on the associated risks and interactions will first be required. Implementing routine testing of a large number of different susceptibility alleles in a substantial set of genes will also require careful deliberation, as it may generate considerable technical and economic burdens for clinical diagnostic services.

Future challenges

These recent advances have underscored the complexity of breast cancer susceptibility, revealing at least three different strata in the genetic architecture of the disease: rare high-penetrance alleles, rare moderate-penetrance alleles and common low-penetrance alleles. It is likely that these categories of susceptibility alleles are germane to many other complex conditions. However, their exploration remains demanding, particularly as the identification of alleles underlying each class requires different strategies and technologies. Moreover, despite the remarkable progress made in the last year, most of the familial risk of breast cancer remains unexplained, highlighting the need for ongoing efforts to expand our view of the emerging landscape of breast cancer susceptibility.

ACKNOWLEDGMENTS

We are grateful to C. Turnbull and R. Scott for their critical reading of the manuscript and helpful comments.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet* **360**, 187–195 (2002).
2. Peto, J. & Mack, T.M. High constant incidence in twins and other relatives of women with breast cancer. *Nat. Genet.* **26**, 411–414 (2000).
3. Hall, J.M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
4. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, *BRCA2*, to chromosome 13q12–13. *Science* **265**, 2088–2090 (1994).
5. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–792 (1995).
6. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
7. Thompson, D. & Easton, D.F. Cancer incidence in *BRCA1* mutation carriers. *J. Natl. Cancer Inst.* **94**, 1358–1365 (2002).
8. The Breast Cancer Linkage Consortium. Cancer risks in *BRCA2* mutation carriers. *J. Natl. Cancer Inst.* **91**, 1310–1316 (1999).
9. Anglian Breast Cancer Study Group. Prevalence and penetrance of *BRCA1* and *BRCA2* mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **83**, 1301–1308 (2000).
10. Peto, J. *et al.* Prevalence of *BRCA1* and *BRCA2* gene mutations in patients with early-onset breast cancer. *J. Natl. Cancer Inst.* **91**, 943–949 (1999).
11. Gudmundsdottir, K. & Ashworth, A. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* **25**, 5864–5874 (2006).
12. Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J. Mammary Gland Biol. Neoplasia* **9**, 221–236 (2004).
13. Antoniou, A.C. & Easton, D.F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
14. Smith, P. *et al.* A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosom. Cancer* **45**, 646–655 (2006).
15. Rahman, N. *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
16. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat. Genet.* **31**, 55–59 (2002).
17. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
18. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
19. Erkko, H. *et al.* A recurrent mutation in *PALB2* in Finnish cancer families. *Nature* **446**, 316–319 (2007).
20. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
21. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
22. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
23. Ahn, J., Urist, M. & Prives, C. The Chk2 protein kinase. *DNA Repair (Amst.)* **3**, 1039–1047 (2004).
24. Shiloh, Y. The ATM-mediated DNA-damage response: taking shape. *Trends Biochem. Sci.* **31**, 402–410 (2006).
25. Peng, M., Litman, R., Jin, Z., Fong, G. & Cantor, S.B. BACH1 is a DNA repair protein supporting BRCA1 damage response. *Oncogene* **25**, 2245–2253 (2006).
26. Xia, B. *et al.* Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol. Cell* **22**, 719–729 (2006).

27. CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am. J. Hum. Genet.* **74**, 1175–1182 (2004).
28. Vahteristo, P. *et al.* A *CHEK2* genetic variant contributing to a substantial fraction of familial breast cancer. *Am. J. Hum. Genet.* **71**, 432–438 (2002).
29. Savitsky, K. *et al.* A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* **268**, 1749–1753 (1995).
30. Thompson, D. *et al.* Cancer risks and mortality in heterozygous *ATM* mutation carriers. *J. Natl. Cancer Inst.* **97**, 813–822 (2005).
31. Howlett, N.G. *et al.* Biallelic inactivation of *BRCA2* in Fanconi anemia. *Science* **297**, 606–609 (2002).
32. Reid, S. *et al.* Biallelic *BRCA2* mutations are associated with multiple malignancies in childhood including familial Wilms tumour. *J. Med. Genet.* **42**, 147–151 (2005).
33. Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group *J. Nat. Genet.* **37**, 934–935 (2005).
34. Reid, S. *et al.* Biallelic mutations in *PALB2* cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* **39**, 162–164 (2007).
35. Xia, B. *et al.* Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. *Nat. Genet.* **39**, 159–161 (2007).
36. Levran, O. *et al.* The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia. *Nat. Genet.* **37**, 931–933 (2005).
37. Seal, S. *et al.* Evaluation of Fanconi anemia genes in familial breast cancer predisposition. *Cancer Res.* **63**, 8596–8599 (2003).
38. van Puijenbroek, M. *et al.* Homozygosity for a CHEK2*1100delC mutation identified in familial colorectal cancer does not lead to a severe clinical phenotype. *J. Pathol.* **206**, 198–204 (2005).
39. Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98**, 1382–1396 (2006).
40. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.* **39**, 352–358 (2007).
41. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
42. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
43. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
44. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
45. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39**, 954–956 (2007).
46. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**, 984–988 (2007).
47. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
48. Pollock, P.M. *et al.* Frequent activating *FGFR2* mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26**, 7158–7162 (2007).
49. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
50. Honrado, E., Benitez, J. & Palacios, J. Histopathology of BRCA1- and BRCA2-associated breast cancer. *Crit. Rev. Oncol. Hematol.* **59**, 27–39 (2006).
51. Heikkinen, K. *et al. RAD50* and *NBS1* are breast cancer susceptibility genes associated with genomic instability. *Carcinogenesis* **27**, 1593–1599 (2006).
52. Tommiska, J. *et al.* Evaluation of *RAD50* in familial breast cancer predisposition. *Int. J. Cancer* **118**, 2911–2916 (2006).
53. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
54. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
55. Domchek, S.M. & Weber, B.L. Clinical management of *BRCA1* and *BRCA2* mutation carriers. *Oncogene* **25**, 5825–5831 (2006).